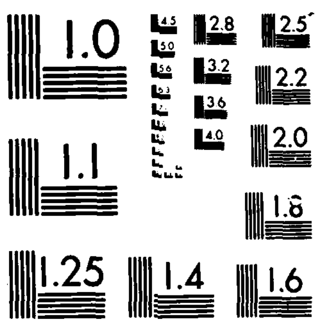
AD-A163 837 MODIFIED KOLMOGOROV-SMIRNOV ANDERSON-DARLING AND 1/2
CRAMER-VON MISES TESTS F..(U) AIR FORCE INST OF TECH
WRIGHT-PATTERSON AFB OH SCHOOL OF ENGI.. J E PORTER
UNCLASSIFIED DEC 85 AFIT/GSO/MA/85D-6 F/G 12/1 NL



AD-A163 837

MODIFIED KOLMOGOROV-SMIRNOV, ANDERSON-DARLING, AND CRAMER-VON MISES TESTS FOR THE PARETO DISTRIBUTION WITH UNKNOWN LOCATION AND SCALE PARAMETERS

THESIS

James E. Porter III
Captain, USAF

AFIT/GSO/MA/85D-6

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

(1)

AFIT/GSO/MA/85D-6

MODIFIED KOLMOGOROV-SMIRNOV,
ANDERSON-DARLING, AND CRAMER-VON MISES
TESTS FOR THE PARETO DISTRIBUTION WITH
UNKNOWN LOCATION AND SCALE PARAMETERS

THESIS

James E. Porter III
Captain, USAF

AFIT/GSO/MA/85D-6

AFIT/GSO/MA/85D-6

MODIFIED KOLMOGOROV-SMIRNOV, ANDERSON-DARLING, AND CRAMER-VON MISES TESTS FOR THE PARETO DISTRIBUTION WITH UNKNOWN LOCATION AND SCALE PARAMETERS

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

in Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Space Operations

James E. Porter III, B.S.

Captain, USAF

December 1985

Approved for public release; distribution unlimited

PREFACE

This thesis develops goodness-of-fit tests for the Pareto distribution by generating critical value tables for the modified Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises statistics. These tables can be used to test whether a set of observed values follows a Pareto distribution when the location and scale parameters are unspecified and must be estimated from the observed sample data. Additionally, the power of each of the three goodness-of-fit tests is studied and compared. Finally, the functional relationship between the critical values and the Pareto shape parameter is determined. Hopefully the material is presented in sufficient detail to be easily understood by those with only a passing knowledge of statistical analysis.

I wish to thank my reader and class advisor, Lieutenant Colonel Joseph Coleman, who guided me throughout my AFIT tour; and especially my thesis advisor, Dr. Albert H. Moore, who maintained my interest in statistical analysis, offered constant encouragement, and suggested the subject of this thesis. I also thank my classmates Majors Dennis Charek and Denny Danielson for their help in debugging the computer programs used in this thesis.

Above all I thank my family, especially my wife Judy, for their love and understanding during my tour at AFIT.

James E. Porter III

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

LIST OF TABLES

AFIT/GSO/MA/85D-6

ABSTRACT

Modified Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Cramer-von Mises (C-VM) critical values are generated for the three-parameter Pareto distribution. The values may be used to test whether a set of observations follows a Pareto distribution when the location and scale parameters are unspecified and thus must be estimated from the sample. A Monte Carlo simulation of 5000 repetitions is used to generate critical values for sample sizes 5(5)30 (i.e., 5 to 30 in increments of 5) and Pareto shape parameters .5(.5)4.0.

A 5000-repetition Monte Carlo investigation is carried out by using 5, 15, and 25 observations from eight alternate distributions to compare the powers of the K-S, A-D, C-VM, and Chi-square tests. The power values of the tests are relatively low for a sample size of five. However, the powers of the modified K-S, A-D, and C-VM tests are considerably better than the Chi-square test at larger sample sizes. Next to the Chi-square test, the A-D test has the lowest power in most cases.

A functional relationship is identified between the modified K-S and C-VM test statistics and the Pareto shape parameter. The critical values are found to be a linear function of the shape parameters between 1.5 and 4.0.

MODIFIED KOLMOGOROV-SMIRNOV, ANDERSON-DARLING, AND CRAMER-VON MISES TESTS FOR THE PARETO DISTRIBUTION WITH UNKNOWN LOCATION AND SCALE PARAMETERS

I. INTRODUCTION

Chapter Overview

This chapter introduces the topic of goodness-of-fit testing and its applications. It states the problem, the research question, and the objectives of the research.

Background

Because the Air Force depends on highly complex weapons systems to perform its missions, factors such as the reliability and maintainability of equipment continue to receive a great deal of emphasis. Of particular importance to the Air Force is the ability to forecast time-to-failure of equipment components and expected maintenance service times.

In studying such phenomena, analysts often face the problem of testing agreement between probability theory and actual observations. When trying to develop a valid statistical model of observed data, the analyst performs four basic steps (5:332):

1. Collect and plot the raw data to develop a histogram (frequency distribution graph).

2. Hypothesize the underlying statistical distribution of the data by comparing the histogram to probability density functions of known distributions.

3. Use the observed data to estimate parameters that characterize the distribution.

4. Test the distributional assumption and parameter estimates for goodness-of-fit. If the hypothesis (that the data follow the assumed distribution) fails, return to step 2 (assume a different distribution) and repeat the process.

Goodness-of-fit tests measure the degree of agreement between the distribution of an observed data sample and a theoretical distribution. Three tests widely used for this purpose are the Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and the Cramer-von Mises (C-VM). Such tests have been developed for several well known distributions, including the normal, exponential, Weibull, gamma, uniform, Laplace, and others (9;19;34;35). However, there are many other distributions which have not been successfully examined for goodness-of-fit when the parameters of the distribution are unknown. One such distribution, which has significant potential for Air Force applications, is known as the Pareto distribution.

The Pareto distribution is an important function in statistical analysis, and several applications have been identified in the fields of economics and operations research. For example, the Pareto distribution has played a major role in investigations concerning the distributions of

city population sizes, natural resources, stock price fluctuations, and oil field locations (28:242). Other studies indicate that the Pareto can be used to model phenomena which may be applicable to Air Force interests, such as time-to-failure of equipment components (16), maintenance service times (22), nuclear fallout dispersion (18), and error clusters in communications circuits (7). Use of the Pareto for such practical applications would be enhanced by an accurate method to test goodness-of-fit of the Pareto distribution.

Problem Statement

A test to determine goodness-of-fit has not been developed for the Pareto distribution when the location and scale parameters are unknown. Such a test would be useful in determining whether a random sample of data taken from an observed phenomenon behaves as the Pareto distribution.

Research Question

How can the existing K-S, A-D, and C-VM tests be modified to produce new goodness-of-fit tests which can be applied to the Pareto distribution when the location and scale parameters are unknown?

Objectives

The objectives of this thesis are to:

1. Generate and document the modified K-S, A-D, and C-VM critical value tables for the Pareto distribution. These tables can be used to test goodness-of-fit when parameters of the distribution are unknown.

2. Compare the powers of the modified K-S, A-D, and C-VM tests to determine which test can best detect a false Pareto distribution hypothesis. The power of a statistical test is the probability of correctly rejecting a false hypothesis.

3. Determine what (if any) functional relationship exists between the shape parameter and the critical values generated for the Pareto function. This relationship can then be used to interpolate critical values corresponding to parameters not found in the generated tables.

Presentation of Research

The report on this thesis effort is presented in seven chapters. In this, the first chapter, the general topic of goodness-of-fit has been introduced and the problem, research question, and objectives have been stated.

Chapter II describes various types of goodness-of-fit tests; explains hypothesis testing and test statistics; and discusses the empirical distribution function.

Chapter III describes applications of the Pareto

distribution; presents its various forms; explores parameter estimation for the Pareto function; and develops the modified K-S, A-D, and C-VM test statistics for the Pareto.

Chapter IV describes the basic principles and specific procedures used to satisfy the research objectives.

Chapter V presents the results of the research effort, including tables of critical values, power comparisons, and regression coefficients.

Chapter VI further discusses the results of the research. Observations are made concerning the tables of critical values, power comparisons, and regression coefficients.

Chapter VII contains conclusions and recommendations based on the conduct and results of the research effort.

Finally, the flow charts and computer programs used to carry out the research are contained in the appendices.

II. GOODNESS-OF-FIT TESTS

Chapter Overview

This chapter briefly reviews the literature to provide a background for goodness-of-fit tests. It also describes hypothesis testing and test statistics as they relate to goodness-of-fit. Finally, it discusses the empirical distribution function and related statistics, including the exact and computational forms of the Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Cramer-von Mises (C-VM) test statistics.

Introduction

Goodness-of-fit tests measure the degree of agreement between the distribution of an observed data sample and a theoretical statistical distribution (13:189). For example, a test for goodness-of-fit may involve examining a random sample from some unknown distribution to test the hypothesis that the underlying distribution is actually a known, specified function (13:345). If such tests indicate a close fit, the hypothesized distribution can then be applied in simulation modeling to predict failure and operational availability rates of Air Force systems and their components.

Background

For years statisticians have attempted to find test statistics whose sampling distributions do not depend on certain parameter values or on the explicit form of the distribution of the population. Such tests are called non-parametric or distribution-free tests (39:68).

Two of the oldest and best known distribution-free tests for goodness-of-fit are the Chi-square and the Kolmogorov-Smirnov (K-S) tests (13:189;47:2). The Chi-square test compares frequencies of the observed data with expected frequencies of the hypothesized distribution. The test is flexible enough to allow some parameters to be estimated from the observed data, but it has some limitations. For example, it is restricted to large sample sizes (1:73). Also, it requires that the data be arbitrarily grouped, which may affect the results (13:357). The K-S test compares the cumulative distribution function (CDF) of the hypothesized distribution against the empirical distribution function (EDF) of the observed data sample. The K-S test can be used for large or small samples; however, it is restricted to distributions which are fully specified (i.e., there can be no unknown parameters that must be estimated from the sample) (13:357). The same limitation applies to two other related methods, the Anderson-Darling (A-D) and the Cramer-von Mises (C-VM) tests (19:204; 47:3-4).

In a significant development, David and Johnson (14) found that if a distribution has only a location and scale parameter, then the K-S and related goodness-of-fit tests are independent of the true parameter values when the parameters are replaced by invariant estimators. The estimators must be invariant in the sense that if each x is transformed by $x \rightarrow ax+b$ then the estimate $T=T(x)$ is similarly transformed by $T \rightarrow aT+b$ (4:4). Therefore, critical values dependent only on sample size and significance level can be generated (54:5). This property also applies to a three-parameter CDF provided the shape parameter is treated as a constant. A more detailed explanation of this principle is included below in the section on "Using Unknown Parameters".

Based on this discovery by David and Johnson, critical value tables for the K-S and related tests have been modified to allow their use in several cases where parameters are estimated from observed data. In a modified test, the form of the test statistic itself remains essentially the same, except that estimates are used in place of exact parameters. However, the critical values for a modified test are considerably different. The critical value tables are no longer the same for all distributions. Instead, they are different for each different hypothesized distribution function. A modified test is still non-parametric or distribution-free because the level of significance is still independent of any untested assumptions regarding the

distribution of the underlying population. In fact, the form of the hypothesized distribution is the hypothesis being tested (13:357).

There are numerous cases for which modified tests have already been developed. For example, Lilliefors developed a modified K-S test for the normal (34) and exponential (35) distributions; Ream (43) developed another set of modified tests for the normal distribution; Woodruff, Moore, and Cortes (53) developed a modified K-S test for the three-parameter Weibull distribution; Bush (9) modified the A-D and C-VM tests to expand the goodness-of-fit tests for the Weibull distribution; Viviano (49) modified the K-S, A-D, and C-VM tests for the gamma distribution; and Yoder (54) developed a modified K-S, A-D, and C-VM test for the logistic distribution. The modified K-S, A-D, and C-VM tests have also been developed for the uniform, normal, Laplace, exponential, and Cauchy distributions (19). Using a different technique, Woodbury (52) too developed a set of modified tests for the uniform distribution.

Hypothesis Testing and Test Statistics

A fundamental concept in statistical testing is the hypothesis test. When studying a given phenomenon, it is often desirable to determine the distribution of the population being studied. In many cases, however, it is not practical to observe the entire population. Instead, a

relatively small sample of the population is usually selected, and observations are made from the small sample.

Hypothesis testing is the process of inferring from a sample whether to "accept" a certain statement (the null hypothesis) about the population from which the sample is drawn. Actually, "acceptance" of the null hypothesis does not imply that the null hypothesis is true, but that there is insufficient evidence from the data sample to reject the hypothesis. The null hypothesis, denoted H_0, is the hypothesis to be tested. The alternative hypothesis, denoted H_1, is equivalent to stating that H_0 is not true (13:75-76).

Another key concept in statistical testing is the test statistic, a function of random variables which is used to help make the decision in a hypothesis test. In order to be useful for data analysis, the test statistic chosen should possess certain desirable properties. Most importantly, the statistic should assign real numbers to points in the sample so that the points are arranged in an order which reflects their ability to distinguish between a true H_0 and a false H_0 (13:77). For example, the test statistic normally assigns larger values to situations that indicate most strongly that H_0 ought to be rejected, while smaller values of the test statistic usually indicate insufficient evidence to reject H_0. In this type of "one-tailed" test, if the value of the test statistic for a given set of data is greater than a certain "critical value", the analyst would reject H_0

(13:77). The critical value is chosen so that when the null hypothesis H_0 is true, the chance of erroneously rejecting H_0 is some specified probability (e.g., .01 or .05) (2:193).

There are two types of errors that can be made in applying the decision criterion. The Type I error results in rejection of H_0 when H_0 is true. The Type II error results in acceptance of H_0 when H_0 is false. The probability of committing a Type I error, denoted by α, is called the level of significance of the test. The probability of a Type II error is denoted β. The power of a statistical test, denoted $1 - \beta$, is the probability of correctly rejecting a false H_0 (13:79).

Statistics Based on the Empirical Distribution Function

One class of test statistic used in goodness-of-fit testing compares an observed sample distribution function and an hypothesized theoretical distribution function. These statistics are based on the empirical distribution function (EDF), and in many cases are easily calculated and competitive in terms of power. The K-S, A-D, and C-VM test statistics are of the EDF type (45:730).

When analyzing phenomenon such as time-to-failure of equipment components, H(x), the actual distribution function of the phenomenon, is rarely known. Often an educated guess of the form of the distribution is made, and the guess is used to approximate the true distribution function. One way

to make a "good guess" is to observe several values from random samples of the phenomenon and construct a graph that can be used to estimate the entire unknown distribution function H(x). One widely used method of constructing such a graph is the empirical distribution function S(x), which equals the fraction of observed values that are less than or equal to x (47:1), i.e.,

$$S(x) = \frac{\text{number of values} \leq x}{\text{total number of values}} \qquad (1)$$

For a sample consisting of n observations, the EDF, which may be denoted $S_n(x)$ to indicate the particular sample size, is a step-shaped function where each step is of height 1/n and occurs only at the sample values. As n becomes larger, $S_n(x)$ should better approximate H(x), provided that H_0 is true. When the n observations are arranged in ascending order, i.e., letting $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the "order statistics" (15:4; 20:70), then $S_n(x)$ is defined (47:1) as:

$$S_n(x) = \begin{cases} 0 & \text{for all } x < x_{(1)} \\ i/n & \text{for } x_{(i)} \leq x < x_{(i+1)},\ i=1,2,\ldots,n-1 \\ 1 & \text{for all } x > x_{(n)} \end{cases} \qquad (2)$$

Like a CDF, $S_n(x)$ is a nondecreasing function that ranges from zero to one in height; however, $S_n(x)$ is determined empirically (from an observed sample), thus its name (13:70).

In a typical test for goodness-of-fit, a random sample from an unknown distribution is examined to test the null hypothesis that the unknown CDF H(x) is in fact a known, specified function F(x), i.e., H_0: H(x) = F(x). The random sample is compared with the hypothesized distribution F(x) in some way to determine whether it is reasonable to conclude that F(x) is the true CDF of the random sample. Using the EDF $S_n(x)$ is one way to compare the random sample with F(x). The fact that $S_n(x)$ is, by definition, the proportion of a random sample less than x implies that it should serve as a good estimate of F(x), which is defined as the probability that the random variable X is less than the value x (47:1). Since the EDF $S_n(x)$ may be useful as an estimator of the hypothesized CDF F(x), then $S_n(x)$ can be compared with F(x) to see if there is close agreement. If the level of agreement is poor, then the null hypothesis is rejected, i.e., the true but unknown CDF H(x) is not the same as the hypothesized function F(x) (13:345).

Based on this approach, the K-S, A-D, and C-VM tests use criteria that measure the discrepancy or "distance" between the hypothesized CDF F(x), which approximates H(x) under H_0, and the EDF $S_n(x)$. The definitions of the three criteria relate to the full range of x, leading to integral forms of the A-D and C-VM test statistics. Conveniently, all three test statistics can be expressed in computational forms in terms of F and S_n at the observed x values (19:204).

Using Unknown Parameters. In their unmodified forms, most popular goodness-of-fit tests based on EDF statistics, including the K-S, C-VM, and A-D tests, are meant to be used only when the null-hypothesized distribution F(x) is fully specified (i.e., when all parameters are known). However, cases are rare in statistical practice when H_0 is completely specified; thus, it is more realistic to have unknown parameters for the null distribution. When unknown parameters are involved, the K-S, C-VM, and A-D tests are no longer distribution-free, so that different critical values will relate to different F(x) in the null hypothesis (19:204). The reason for this is that the distributions of these and other EDF statistics depend on the sample size n and also on the values of the unknown parameters (47:4).

The K-S, C-VM, and A-D tests depend on the probability integral transformation described by David and Johnson (14). This transformation, when applied to a random sample from a distribution of specified parameters, produces ordered values from a uniform distribution over the interval from 0 to 1. These values are then used to calculate the EDF test statistic. As a result, the EDF statistic becomes a function of ordered uniform random variables. However, when parameters are unknown and must be estimated from the sample, the transformation fails to produce ordered uniform random variables (47:4). Unless appropriately modified, therefore, any EDF tests based on this transformation will generally be

restricted to cases where all parameters are specified.

An important exception occurs if the unknown parameters are location and scale only. David and Johnson (14) showed that if a distribution can be completely specified by a single parameter for location and a single parameter for scale, then goodness-of-fit tests based on the probability integral transformation are independent of the true parameter values when invariant estimators are used (38:384).

Fortunately, the Pareto distribution can be completely specified by a single location and a single scale parameter (28:239). The three-parameter form of the Pareto, presented in the next chapter, can be expressed in terms of a single location and scale parameter by treating the shape parameter as a known constant. *Thus, the value of each EDF test statistic for the Pareto will depend only on the sample size and significance level, but not on the exact values of the unknown parameters* (35:387). As a result, rather than having to produce a separate set of critical value tables for each set of location and scale parameters, only one set of tables is needed for each shape parameter and each sample size n. It is this principle, coupled with the fact that the Pareto possesses the necessary location and scale property, that allows the generation of valid critical value tables for the Pareto distribution (47:5).

To accomplish this goal, the existing (unmodified) K-S, A-D, and C-VM test statistics can be modified using an

invariant estimator; but first, the unmodified statistics are discussed in the following sections.

The Kolmogorov-Smirnov Statistic. The K-S statistic in its unmodified form is especially useful when sample sizes are small and when no parameters are estimated from the data. Often it is a more powerful test than the Chi-square for any sample size (34:399; 39:76). However, when parameter estimates must be made from the sample, the Chi-square test is easily modified by reducing the number of degrees of freedom, whereas the existing K-S critical values are overly conservative and must be modified using Monte Carlo techniques (5:357). In this context, the term "conservative" means that the critical values are too large so that the actual level of significance is smaller than the stated level of significance (13:90).

The K-S test statistic (36:259-260; 5:270; 19:204) is the largest (denoted "sup" for supremum) vertical distance between the completely specified hypothesized CDF F(x) and the observed EDF $S_n(x)$. Therefore, the test statistic is expressed as:

$$D = \sup_x |F(x) - S_n(x)| \qquad (3)$$

which is equivalent to the computational form given by

$$D = \max\ (D^+,\ D^-) \qquad (4)$$

H_0 is rejected if D exceeds a corresponding critical value (13:358).

If there are n observations, $x_{(i)}$ is the i-th smallest observation, and $z_i = F(x_{(i)})$ then (39:69):

$$D^+ = \sup_{1 \leq i \leq n} [(i/n)-z_i] \quad \text{and} \quad D^- = \sup_{1 \leq i \leq n} [z_i-(i-1)/n] \qquad (5)$$

Thus the K-S statistic is the larger of these two values.

<u>The</u> <u>Cramer-von Mises</u> <u>Statistic</u>. Another way to measure the discrepancy between the hypothesized CDF F(x) and the observed EDF $S_n(x)$ is to use statistics of the Cramer-von Mises family, based on the squared integral of the difference between the EDF and the distribution tested (47:2). One such statistic is the C-VM statistic itself (46:357):

$$W^2 = n\int_{-\infty}^{\infty}[S_n(x)-F(x)]^2 dF(x) \qquad (6)$$

which in computational form is (3:766; 45:731):

$$W^2 = [1/(12n)] + \sum_{j=1}^{n}[z_j - (2j-1)/2n]^2 \qquad (7)$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are n ordered observations from the sample and $z_j = F(x_{(j)})$ for $j=1,2,\ldots,n$.

The Anderson-Darling Statistic. Another member of the Cramer-von Mises family is the A-D statistic. To allow more flexibility in goodness-of-fit tests, Anderson and Darling (2:194) introduced the technique of incorporating a weight function into the K-S and C-VM test statistics. The result is still another method of testing the hypothesis that n observations have been drawn from a population with specified distribution function F(x).

Anderson and Darling (3:767) suggested using a nonnegative weight function, here denoted $\theta(u)$, chosen by the analyst to accentuate the values of $S_n(x) - F(x)$ in those areas where the test is desired to have greater sensitivity. This weight function serves to counteract the fact that the discrepancy between $S_n(x)$ and $F(x)$ becomes smaller in the tails, since each approaches 0 and 1 at the extremes (47:2). They found that choosing the weight function θ in the form of $\theta(u) = 1/[u(1-u)]$ has the effect of heavily weighting the discrepancy in the tails of the two distributions. The resulting A-D test statistic (2:193; 46:357) is:

$$A^2 = n\int_{-\infty}^{\infty}[S_n(x)-F(x)]^2\theta[F(x)]dF(x) \tag{8}$$

$$\text{where} \quad \theta[F(x)] = [F(x) \cdot (1-F(x))]^{-1}$$

Thus the C-VM statistic may be considered a special case of the A-D statistic where $\theta[F(x)] = 1$.

In computational form the A-D statistic is (3:765):

$$A^2 = -n - (1/n) \sum_{j=1}^{n} (2j-1)[\ln z_j + \ln(1-z_{n+1-j})] \qquad (9)$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are n ordered observations from the sample and $z_j = F(x_{(j)})$ for $j=1,2,\ldots,n$.

The A-D statistic is designed to be used when the analyst wants the test to have good power against alternatives in which F(x) and H(x), the true distribution, disagree near the tails of F(x), and is willing to sacrifice power against alternatives in which they disagree near the median of F(x) (3:767). Thus, the A-D statistic is used when the analyst wants to reject H_0 if H(x) differs greatly from F(x), and especially if the difference is in the tails.

Chapter Summary

The K-S, A-D, and C-VM tests are non-parametric tests of goodness-of-fit which offer advantages over the older Chi-square test. In their usual forms, the K-S, A-D, and C-VM tests are restricted to distributions which are fully specified. However, when location and scale parameters are replaced by invariant estimators, the three tests can be modified to produce valid critical values for a given distribution. Hypothesis testing and test statistics are two statistical concepts which can be used to modify the existing tests for the Pareto distribution, which is discussed in detail in the next chapter.

III. THE PARETO DISTRIBUTION

Chapter Overview

This chapter reviews the history and application of the Pareto Law; presents the Pareto distribution and its three parameters; explores parameter estimation for the Pareto function; and develops the modified Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Cramer-von Mises (C-VM) test statistics for the Pareto distribution.

History and Application

Origin. The Pareto distribution is an important function in statistical analysis. It is named after Vilfredo Pareto (1848-1923), a Swiss professor of economics who conducted the first extensive statistical study of the distribution of incomes. His analysis of nineteenth century income in various countries led to the development of his first law:

> . . . if x signify [sic] a given income and N the number of persons with incomes exceeding x, and if a curve be drawn, of which the ordinates are logarithms of x and the abscissae logarithms of N, this curve, for all the countries examined, is approximately a straight line . . . This means that, if the number of incomes greater than x is equal to N, the number greater than mx is equal to $N/m^{1.5}$, whatever the value of m may be. Thus the scheme of income distribution is everywhere the same [42:647].

Therefore, "the logarithm of the percentage of units with an income greater than some value is a linear function of that value with negative slope, provided that this value is greater than an appropriate positive number" (32:6). This is known as the "strong" form of the Pareto Law, with functional form given by equation (11) below. The "weak" form of the law pertains to the asymptotic nature of a distribution's tail and implies that if log $[1-F_X(x)]$ is plotted against log x, then the resulting curve should be asymptotic to a line with slope $-c$ as x gets larger (32:6; 28:245).

Early Applications. Since the early days of its formulation, the Pareto Law and its related distribution functions have been examined primarily for potential applications in economics and operations research.

Based on his statistical observations, Pareto believed that any influence that causes an increase in the national income overall must also increase the income of the poor: "We cannot be confronted with any proposal the adoption of which would both make the dividend larger and the share of the poor smaller, or vice versa" (42:648). Pareto also believed his law to be universally inevitable, regardless of economic, social, and political conditions. Economists have since identified flaws (11:609; 17:171) in the Pareto Law to the extent that for several years the Pareto distribution became disreputable (28:233; 7:235) as an economic predictor:

> The general defence of "Pareto's Law" as a law of even limited necessity rapidly crumbles. His statistics warrant no inference as to the effect on distribution of the introduction of any cause that is not already present . . . This consideration is really fatal; and Pareto is driven, in effect, to abandon the whole claim [42:654].

Nevertheless, more recent studies have shown the Pareto distribution can be very useful.

Recent Applications. Several more recent studies have revived interest in the Pareto distribution by demonstrating that it can be used to model or predict numerous empirical phenomena. For example, the Pareto distribution has played a major role in investigations concerning city population size, resources, stock price fluctuations, and oil fields (28:242). The Pareto has also been used to describe property values, inheritance, business mortality, worker migration, consumer prices, and effects of underreported income (32:7; 51).

Fisk (17:171, 174-175) showed that in some cases the Pareto distribution offers an improvement over the lognormal distribution, especially at the extremities (tails) of the distribution. Steindl (44:187-246) cited several examples of empirical economic data which follow the Pareto distribution, including the distribution of wealth, jobs by basic salary, the growth rate of firms and corporations, and several others. He also reaffirmed the Pareto Law's usefulness in economic theory:

> Empirical laws are rare in economics, and the most obvious instance of such laws is the regular pattern of certain statistical distributions, such as the distribution of persons according to income or of business firms according to sales. A good many of these distributions conform to the so-called law of Pareto, i.e. the number of firms (for example) with sales in excess of X, plotted against X on logarithmic paper, is a straight line . . . The Pareto distribution is encountered in many fields and often the fit is very good [44:11].

Air Force Applications. Other studies have shown that the Pareto can be used to model phenomena which may be applicable to Air Force interests, such as time-to-failure of equipment, maintenance service times, nuclear fallout particles, and error clusters in communication circuits.

For example, Davis and Feldstein (16:299) showed the Pareto can be used to model survival data based on a population of items whose times-to-failure from a well defined origin are being observed. If each member of the population has a constant hazard rate based on a two-parameter gamma distribution, then the time-to-failure for the population is the Pareto type II of equation (13). Further, in some cases the Pareto competes with the Weibull distribution as a model for failure times of components. Like the Weibull, the generalized Pareto includes the exponential, and can therefore be used to test departures from the exponential (16:305-306).

Kaminsky and Nelson (30) showed how the Pareto distribution can be used in situations involving life testing,

reliability, and replacement policy. Specifically, they showed how to use the Pareto to predict the time of future failures from times of early failures in the same sample. They found, for example, that if items are put into service simultaneously, and it becomes necessary to begin replacing them when a certain percentage remain functional, then it is possible to predict the replacement time of future failures from the early failure times. In another example, "if n items form an n-component parallel system, then we can predict the time of system failure . . ." (30:145).

The Pareto distribution can also be of use in modeling queuing systems in which equipment maintenance service times are conditioned upon a random parameter. Harris (22:307) showed that if the conditional service distribution is exponential and the random parameter has a gamma density, then the resultant service times follow the Pareto distribution. Further, if a system consists of components which have exponentially distributed times-to-failure with a gamma parameter density, then the unconditional times to failure would follow the Pareto distribution (22:312). Harris also used the Pareto to develop a model which provides a means of obtaining measures of effectiveness of a large scale and complicated queuing process (22:308-309).

Freiling showed that the Pareto distribution, in the form of equation (10) with c = 3, can be used to model mass sizes of nuclear fallout particles (18:4). In addition, he

compared the usefulness of the Pareto and lognormal distributions in modeling the size distribution of particle mass in the fallout from land-surface bursts. For this specific application, Freiling found close similarities between the two distributions: "The agreement is such that if one curve is correct, the other will never be proved wrong . . . Thus it appears that the differences between the two approaches are trivial" (18:12). He concluded his study by noting that, in the case of nuclear airburst debris, the lognormal distribution has the advantage of having an "observationally confirmed theoretical basis." If the observational data is truncated, however, the Pareto distribution has the advantage of simplifying calculations of particle surface distribution.

In a study of error clusters in communication circuits, Berger and Mandelbrot (7:224) revealed still another application of the Pareto distribution. They proposed a new mathematical model to describe the distribution of the occurence of errors in data transmission over telephone lines. They found that the statistics of communications errors can be described in terms of an error probability depending solely on the time elapsed since the last occurrence of an error. Further, they discovered that the distribution of inter-error intervals closely approximates the Pareto distribution of exponent less than one. As a result, the relative number of errors tend to zero as message lengths increase.

The Pareto Function

Pareto's Law in its original form can be expressed as $N = Ax^{-c}$ where A and c are parameters which characterize the function and N is the number of people having income of at least x. In a form more commonly used in statistical analysis, Pareto's Law becomes the Pareto distribution:

$$P(x) = Pr[X \geq x] = (k/x)^c \quad \text{for } k,c > 0;\ x \geq k \tag{10}$$

where P(x) is the probability that the value of a random variable X (e.g., income) is at least x, k is a lower bound on X (e.g., some minimum income), and c characterizes the shape of the graph of the distribution (28:233-234).

Accumulated probabilities over the range of values of x are given by the corresponding cumulative distribution function (CDF) of X, also known as the "Pareto distribution of the first kind" (28:234) or the "strong" Pareto law (32:50):

$$F_X(x) = 1 - (k/x)^c \quad \text{for } k,c > 0;\ x \geq k \tag{11}$$

The corresponding Pareto probability density function is:

$$p_X(x) = ck^c/x^{c+1} = (c/k)(k/x)^{c+1} \quad \text{for } c > 0;\ x \geq k > 0 \tag{12}$$

Pareto proposed two other forms of the distribution. The "Pareto distribution of the second kind" (also called the Pareto Type II or the Lomax distribution), is:

$$F_X(x) = 1 - K_1/[(x+C)^c] \tag{13}$$

The third form proposed by Pareto, the "Pareto distribution of the third kind" (or Pareto Type III), has the CDF:

$$F_X(x) = 1 - k_2e^{-bx}/[(x+C)^c] \tag{14}$$

which reduces to the Type II form when b = 0.

The basic difference between these various forms is in the number of parameters. The Pareto distribution of the first kind, equation (11), represents the "usual formulation" of the function and is the one most commonly found in the literature. However, the fact that it consists of only two parameters (i.e., c and k) may limit its usefulness in general applications. Hastings and Peacock (26) regard three types of parameters as basic to any distribution function. These three parameters are the location, scale, and shape parameters, which they denote as a, b, and c respectively. The location parameter (a) represents "the abscissa of a location point (usually the lower or midpoint) of the range of the variate." The scale parameter (b) is "a parameter which determines the scale of measurement of the fractile x".

Finally, the shape parameter (c) "determines the shape . . . of the distribution function within a family of shapes associated with a specified type of variate" (26:20).

Kulldorff and Vannman (33:218) introduced a more general form of the CDF than the two-parameter form shown in equation (11). By using the parameter notation of Hastings and Peacock, and the functional form of Kulldorff and Vannman, the generalized (three-parameter) form of the Pareto distribution is illustrated in Figure 1 and can be written as:

$$F(x) = 1 - [1 + (x-a)/b]^{-c} \quad \text{for } x \geq a; \quad b,c > 0 \qquad (15)$$

where again a is location, b is scale, and c is shape.

In the special case when a = b, if we let k = a = b as in Figure 2, then from equation (15):

$$\begin{aligned} F(x) &= 1 - [1 + (x-a)/b]^{-c} = 1 - [1 + (x-k)/k]^{-c} \\ &= 1 - [1 + (x/k) - (k/k)]^{-c} = 1 - (1 + x/k - 1)^{-c} \\ &= 1 - (x/k)^{-c} = 1 - (k/x)^{c} \end{aligned}$$

where $k,b,c > 0$ and $x \geq k = a$. The last expression is the "usual formulation" given by equation (11).

Another form commonly found in the literature (26:102; 51:1) is the one-parameter form (Figures 3 and 4) given by:

$$F(x) = 1 - x^{-c} \text{ for } x \geq 1; \; c > 0 \qquad (16)$$

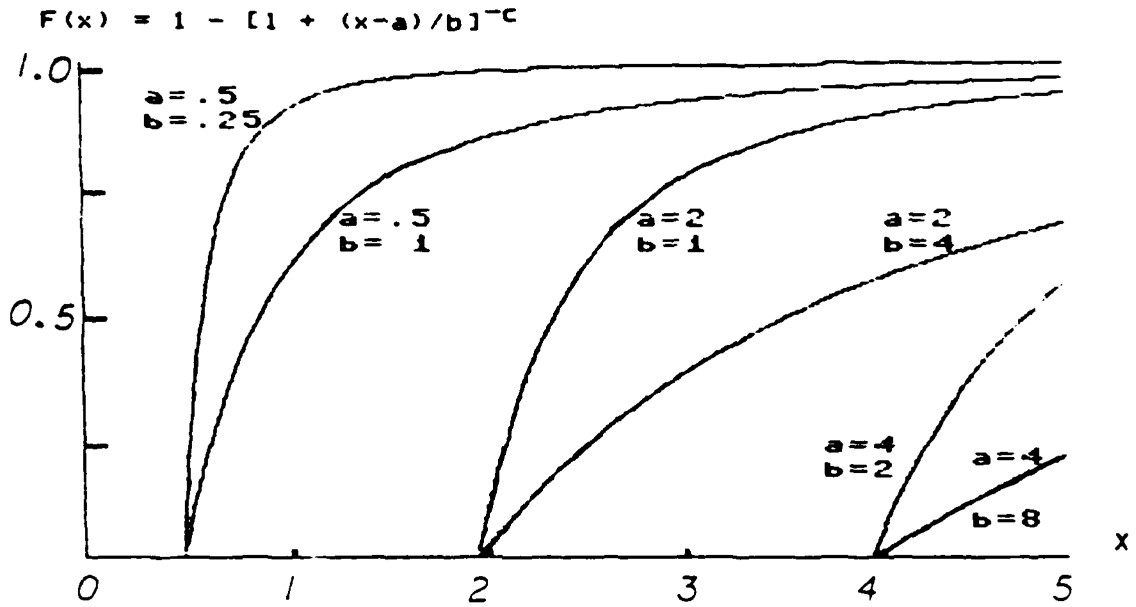


Fig 1. Three-Parameter Pareto Curves (Eqn 15) for Several Values of Location a and Scale b with Shape c = 2.

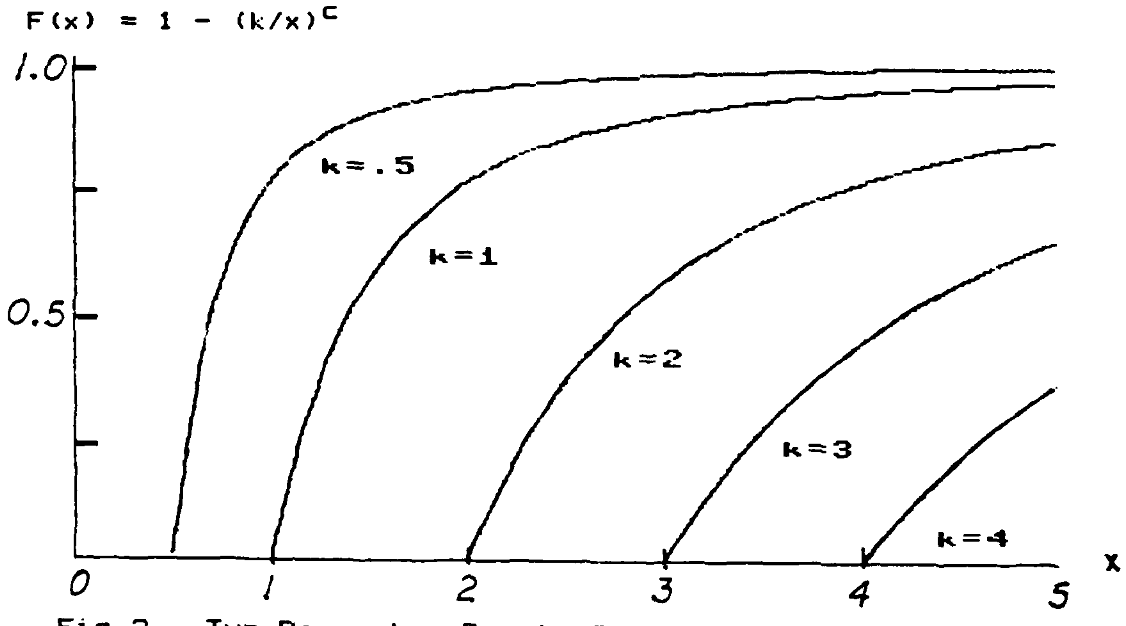


Fig 2. Two-Parameter Pareto Curves (Eqn 11) for Several Values of k with Shape c = 2 and a = b = k.

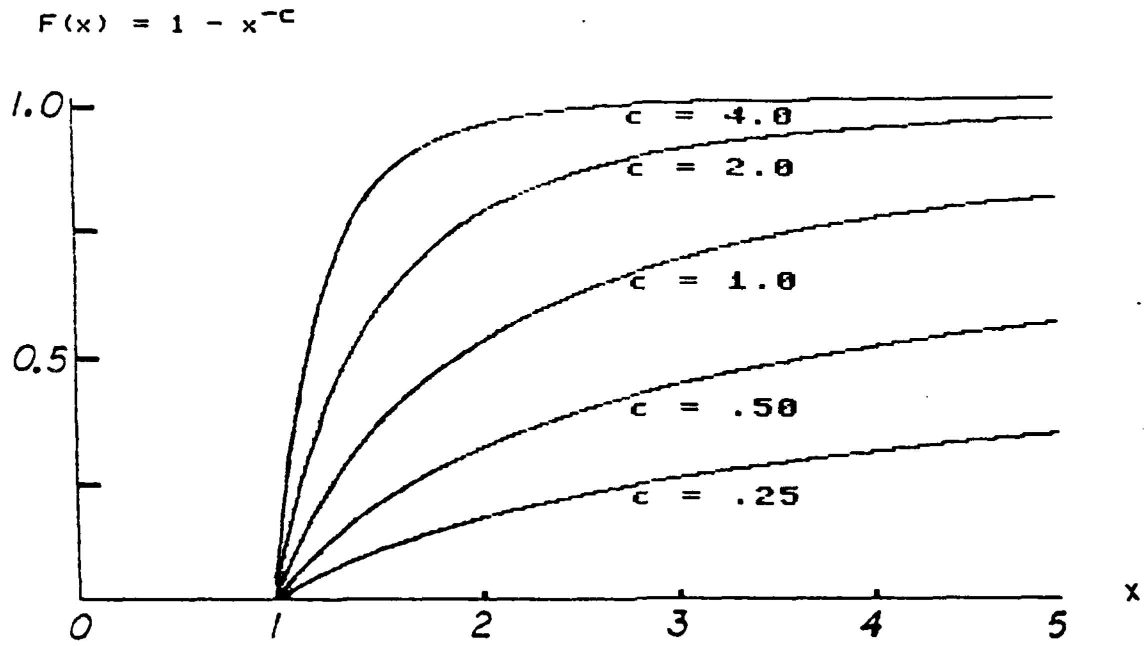


Fig 3. One-Parameter Pareto Curves (Eqn 16) for Several Values of Shape c with k = a = b = 1.

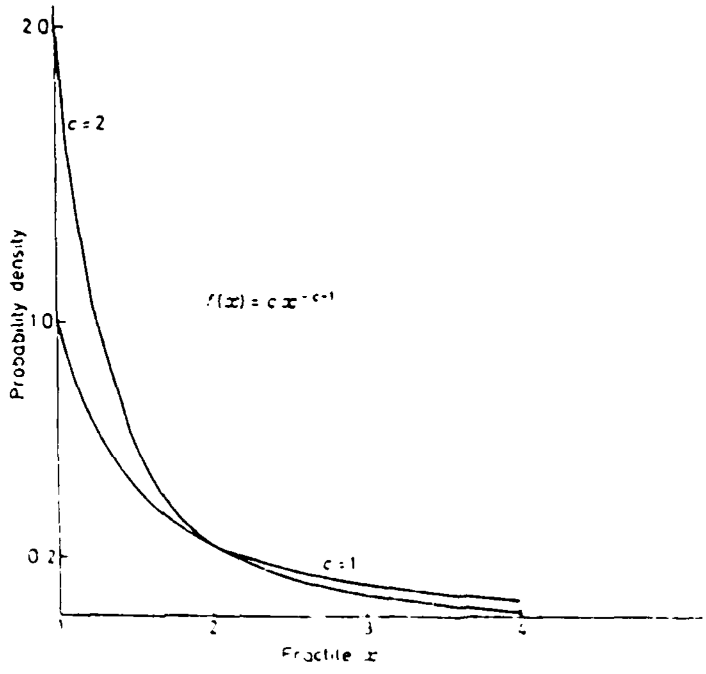


Fig 4. Probability Density (Eqn 12) of the One-Parameter Pareto with k = 1 (Reprinted from 26:103).

Equation (16) is simply a special case of (15) found by setting $a = b = 1$. As such, it represents the least general form of the Pareto distribution.

The greater generality inherent in the three-parameter form, equation (15), allows the Pareto distribution to be more useful in practical applications. For example, in some situations the random variable represented by x may be positive by its very nature, making the assumption $a = 0$ more realistic than $a = b$ (33:218). In the special case where $a = 0$, the three-parameter Pareto distribution becomes:

$$
\begin{aligned}
F(x) &= 1 - [1 + (x-a)/b]^{-c} = 1 - (1 + x/b)^{-c} \\
&= 1 - (b/b + x/b)^{-c} = 1 - [(x+b)/b]^{-c} \\
&= 1 - [b/(x+b)]^{c} = 1 - b^{c}/[(x+b)^{c}]
\end{aligned}
$$

This last expression can be written as equation (13) by simply setting $b^c = K_1$ and $b = C$.

Therefore, equations (11), (13), and (16) each represent special cases of the three-parameter form given by equation (15). Since (15) is a more general and hence more useful form of the Pareto distribution, this thesis uses the functional form in (15) to develop the goodness-of-fit tests for the Pareto distribution. Selecting the more general form as a basis for the test statistics will ensure the widest possible application of the goodness-of-fit tests.

Parameter Estimation

As explained in Chapter II, the development of modified Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises tests depends on the use of an invariant estimator for the unspecified location and scale parameters (38:384). This section begins by briefly examining several published studies on various estimation techniques for Pareto distributions. It concludes by discussing the best linear unbiased estimator (BLUE), which is the invariant estimator used in this thesis.

Various Estimators. The two methods of invariant estimation most commonly used in modified goodness-of-fit tests are the maximum likelihood estimator (MLE) and the best linear unbiased estimator (BLUE). Various techniques for estimating the parameters of the Pareto distribution can be found in the literature. However, as Kulldorff and Vannman (33:218) point out, few studies consider the general three-parameter form of equation (15). Instead, most studies consider only "special cases", such as $a = b$, corresponding to equations (11) and (12).

Numerous examples of "special case" estimators can be cited. Moore and Harter (41; 23:69,86) developed a biased, single-order-statistic MLE for the Pareto shape parameter when location is specified. Harris (22:308, 310-311) considered estimation for the two-parameter form given by

equation (12): "As a first try, we can appeal to the techniques of maximum likelihood estimation. However, this particular method does not yield sufficiently simple equations (for even numerical methods)" (22:310). As a result, Harris resorted to the method of moments instead. Johnson and Kotz (28:234-240) presented MLEs for the two-parameter form in equation (11), as well as several other estimation techniques. Hastings and Peacock (26:102) gave the MLE for the one-parameter form of equation (16). In his dissertation, Koutrouvelis (32:97-115) attempted to estimate the parameters of the upper tail of Pareto distributions, but found it too difficult to calculate the Pareto MLEs, even with a computer. Instead, he developed a new method of estimating parameters based on the asymptotic theory of quantiles using only data consisting of sample values greater than some specified value. Wingo (50) wrote a FORTRAN program to calculate the MLEs from a reduced log-likelihood function for the two-parameter form in equation (12). Davis and Feldstein (16:299-300, 305) developed MLEs from progressively censored data for the Pareto Type III, equation (14). Bell, Ahmad, Park, and Lui (6:4-7) presented the MLEs, the minimum variance unbiased estimators (MVUEs), and the minimal sufficient statistic (MSS) for the two-parameter form, equation (11). Several other estimation studies are cited by Koutrouvelis (32:55) and Johnson and Kotz (28:235-240). Unfortunately, none of these studies provide

the invariant estimators of the three-parameter form in equation (15) as needed for this thesis.

Parameter estimation for the general case given by equation (15) went virtually ignored until Kulldorff and Vannman (33) derived the BLUEs of the unknown parameters on the basis of a complete Pareto sample with shape $c > 2$. In a follow-up paper, Vannman (48) derived the BLUEs for shape $c \leq 2$. Later, Kaminsky (29:7-8, 12-14) and Kaminsky and Nelson (30:148) extended the work of Kulldorff and Vannman by deriving, for equation (15), the best linear unbiased predictors of future observations from censored samples. Most recently, Charek (12) examined minimum distance estimation for the three-parameter Pareto.

Best Linear Unbiased Estimator (BLUE). The BLUE derives its name from its main properties as an estimator. It is a "linear" estimator because it can be expressed as a linear function of a random sample. It is "unbiased" because its bias term is zero; and the expected value of the estimator is equal to the true parameter value. It is considered the "best" estimator because it has the minimum variance among all other linear unbiased estimators (27:227). However, for the purposes of this thesis, the most important property of the BLUE is invariance under transformation of parameters.

The BLUE is a subset of a larger class of estimators

known as least-squares estimators. In general, least squares estimators do not possess the invariance property. However, when a least-squares estimator is also a linear function, then the invariance property holds (40:349-350). Therefore, in addition to its other properties, the BLUE is also an invariant estimator. It is this property of invariance under parameter transformations that allowed, for example, Green and Hegazy (19:205) and Woodbury (52) to use the BLUE in producing modified goodness-of-fit tests based on the findings of David and Johnson (14).

Intuitively, the property of invariance implies, for example, that if a parameter θ is estimated, and θ^2 is also estimated from the same data, then the estimate of θ^2 should be the square of the estimate of θ (37:434). Generally, the invariance property requires that if $f(\theta)$ is a single valued function of a parameter θ, and $\hat{\theta}$ is the BLUE of θ, then $f(\hat{\theta})$ is the BLUE of $f(\theta)$, i.e., $f(\hat{\theta}) = \widehat{f(\theta)}$ (8:94).

The studies by Kulldorff and Vannman (33; 48) derived the BLUEs of equation (15) for b when a and c are known; for a when b and c are known; and for a and b when c is known. The last case, which corresponds to invariant estimation of location and scale when shape is known, is used in this thesis to develop the modified K-S, A-D, and C-VM tests. The next two subsections use the findings of Kuldorff and Vannman to derive computational forms of the BLUEs for the Pareto location and scale parameters, assuming shape is known.

__BLUEs for Shape $c > 2$.__ For the case where $c > 2$, Kulldorff and Vannman (33:224-226) found that the BLUEs of location a and scale b can be written in terms of the specified shape parameter c and the order statistics (15:4) $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ where $x_{(1)}$ is the smallest and $x_{(n)}$ the largest value in the observed random sample of size n. Thus the BLUEs for a and b are, respectively:

$$\hat{a} = x_{(1)} - Y/[(nc-1)(nc-2)-ncD] \tag{17}$$

$$\begin{aligned}\hat{b} &= Y(nc-1) \,/\, [(nc-1)(nc-2)-ncD] \\ &= [x_{(1)}-\hat{a}](nc-1)\end{aligned} \tag{18}$$

In the special case when it is known that a = b, as in equation (11), the BLUE reduces to:

$$\hat{k} = [1 - 1/(nc)]x_{(1)} \tag{19}$$

However, before equations (17) and (18) can be used to find the BLUEs for the general case, the following terms must first be calculated:

$$B_i = \frac{\Gamma(n-i+1)\,\Gamma(n+1-2/c)}{\Gamma(n-i+1-2/c)\,\Gamma(n+1)} \quad \text{for } i = 1,2,\cdots,n \tag{20}$$

$$D = (c+1)\sum_{i=1}^{n-1} B_i + (c-1)B_n \tag{21}$$

$$Y = (c+1) \sum_{i=1}^{n-1} B_i \, x_{(i)} + (c-1)B_n \, x_{(n)} - Dx_{(1)} \qquad (22)$$

After these values are calculated, they can be substituted into equations (17) and (18) to find the BLUEs $\hat{a}$ and $\hat{b}$.

From equations (17) to (22), it is obvious that the use of the BLUEs $\hat{a}$ and $\hat{b}$ involves the computation of all the coefficients B_i for $i = 1,2,\cdots,n$. Therefore, in order to derive a computational form of the BLUEs, the first task is to simplify equation (20). Each B_i is the ratio of a product of gamma functions. Banks and Carson (5) note that "the gamma function can be thought of as a generalization of the factorial notion which applies to all positive numbers, not just integers" (5:144). For any real $m > 0$:

$$\Gamma(m) = (m-1)\,\Gamma(m-1) \qquad (23)$$

By definition $\Gamma(1) = 1$, so that whenever m is an integer, equation (23) becomes:

$$\Gamma(m) = (m-1)! \qquad (24)$$

Applying these gamma definitions in equation (20) reveals:

$$B_1 = \frac{\Gamma(n-1+1)\,\Gamma(n+1-2/c)}{\Gamma(n-1+1-2/c)\,\Gamma(n+1)} = \frac{\Gamma(n)\,\Gamma(n+1-2/c)}{\Gamma(n-2/c)\,\Gamma(n+1)}$$

$$= \frac{(n-1)!\ (n-2/c)\,\Gamma(n-2/c)}{n(n-1)!\,\Gamma(n-2/c)} = \frac{n-2/c}{n}$$

$$= 1 - 2/(cn) \tag{25}$$

Similarly, B_2 is found from equation (20) as follows:

$$B_2 = \frac{\Gamma(n-2+1)\,\Gamma(n+1-2/c)}{\Gamma(n-2+1-2/c)\,\Gamma(n+1)} = \frac{\Gamma(n-1)\,\Gamma(n+1-2/c)}{\Gamma(n-1-2/c)\,\Gamma(n+1)}$$

$$= \frac{(n-2)!\ (n-2/c)\,\Gamma(n-2/c)}{\Gamma(n-1-2/c)\ n!}$$

$$= \frac{(n-2)!\ (n-2/c)(n-1-2/c)\,\Gamma(n-1-2/c)}{n(n-1)(n-2)!\,\Gamma(n-1-2/c)}$$

$$= \frac{(n-2/c)(n-1-2/c)}{n(n-1)}$$

$$= [1 - 2/(cn)]\ [1 - 2/c(n-1)] \tag{26}$$

Continuing in this manner, it turns out that:

$$B_n = [1 - 2/(cn)][1 - 2/c(n-1)]\cdots[1 - 2/c(1)] \tag{27}$$

The calculations can be simplified as follows:

Let $g_1 = 2/\ (cn)$, $g_2 = 2/\ [c(n-1)]$, $\cdots$, $g_n = 2/c$.

Also let $b_1 = 1-g_1$, $b_2 = 1-g_2$, $\cdots$, $b_n = 1-g_n$.

Then $B_1 = b_1$, $B_2 = b_1 b_2$, $\cdots$, $B_n = b_1 b_2 \cdots b_n$.
In general, then, each B_i can be expressed in computational form as:

$$B_i = \prod_{j=1}^{i} b_j \qquad (28)$$

where $b_j = 1 - g_j$ and $g_j = 2/\,c(n-j+1)$ for $j = 1,2,\cdots,i$.
From these results, if we let $B_0 = 1$, then another way to write B_i is (48:705):

$$B_i = [1 - 2/\,c(n-i+1)]\, B_{i-1} \quad \text{for } i = 1,2,\cdots,n \qquad (29)$$

As mentioned earlier, once all of the B_i are computed from equation (28) or (29), then D and Y can be computed from equations (21) and (22). Finally, these values for B_i, D, and Y are substituted into equations (17) and (18) to find the BLUEs $\hat{a}$ and $\hat{b}$.

BLUEs for Shape $c \leq 2$. For the case where $c \leq 2$, the variance of the Pareto distribution does not exist, so a different approach must be used to derive the BLUEs. In this case, Vannman (48:706-707) found that the BLUEs of location a and scale b can still be found provided that shape c satisfies $2/n < c \leq 2$, where again n is the sample size. Here the BLUEs a_k^* and b_k^* are based on the first k order statistics only, where k is chosen so that $2 \leq k < n+1-2/c$:

$$a_k^* = x_{(1)} - b_k^*/(nc-1) \tag{30}$$

and

$$b_k^* = (1/U_k) \{ (c+1) \sum_{i=1}^{k-1} B_i \; x_{(i)}$$

$$+ [(n-k+1)c - 1] \, B_k \, x_{(k)}$$

$$- [(nc-1)/(nc)] \, (nc-2-U_k) \, x_{(1)} \} \tag{31}$$

where

$$U_k = \frac{(nc-2)(nc-c-2) - nc[(n-k)c - 2] \, B_k}{(nc-1)(c+2)} \tag{32}$$

Whenever possible, k should be chosen to achieve highest efficiency, which occurs when $k = n - [2/c]$, where "[2/c]" denotes the integer portion of 2/c. Vannman (48:707) also points out that in the case where 2/c is an integer, and k is selected for highest efficiency so that $k = n - 2/c$, then equation (31) can be simplified to:

$$b_k^* = \frac{(c+1)(c+2)(nc-1)}{(nc-2)(nc-c-2)} \left[\sum_{i=1}^{n-2/c} B_i \; x_{(i)} - \frac{nc-2}{c+2} \, x_{(1)} \right] \tag{33}$$

By substituting this result for b_k^* in equation (30), the BLUE for a, based on the first n-2/c order statistics, can be written in the following computational form:

$$a_k^* = x_{(1)} - \frac{(c+1)(c+2)}{(nc-2)(nc-c-2)} \left[\sum_{i=1}^{n-2/c} B_i \, x_{(i)} - \frac{nc-2}{c+2} x_{(1)} \right]$$

(34)

Once a_k^* has been computed, it is easy to use equation (30) to find a computational form of the BLUE for b:

$$b_k^* = b_{n-2/c}^* = (nc-1) \; (x_{(1)} - a_k^*) \qquad (35)$$

Equations (34) and (35) give the BLUEs for location a and scale b provided all of the following conditions apply:

1) shape parameter c is specified
2) $2/n < c \leq 2$
3) $2/c$ is an integer

When sample size n = 5, 10, 15, 20, 25, or 30, then all three of these conditions hold for shape parameter c = .5, 1, or 2. Therefore, for these values of n and c, it appears that equations (34) and (35) apply. There is, however, one important exception. As explained earlier, k must be chosen so that $2 \leq k < n+1-2/c$. In the case where n = 5 and c = .5, notice that $n+1-2/c = 2$. Thus k cannot be selected as before, since it would need to satisfy $2 \leq k < 2$, which is not possible. As a result, the above equations fail to provide BLUEs for the special case c = .5 and n = 5; thus, when c = .5, this thesis will use n = 6 instead of n = 5.

As explained in the next chapter, this thesis uses sample sizes of n = 5, 10, 15, 20, 25, and 30, with shape

parameters of c = .5, 1, 1.5, 2, 2.5, 3, 3.5, and 4. The preceding subsection presented the BLUEs to be used for c = 2.5, 3, 3.5, and 4. This subsection has thus far shown that equations (34) and (35) provide the BLUEs for c = .5, 1, and 2, except for the special case c = .5 and n = 5. The one remaining case to be addressed is when c = 1.5.

When the shape parameter c = 1.5, equations (34) and (35) do not apply since condition 3) fails to hold, i.e., 2/c is not an integer. To ensure highest efficiency, k is selected so that $k = n - [2/c]$, where "[2/c]" denotes the integer portion of 2/c. Thus:

$$k = n - [2/c] = n - [1.333] = n - 1 \tag{36}$$

According to Vannman (48:707), substituting this value of k into equations (30) to (32) gives the desired BLUEs:

$$a_k^* = a_{n-1}^* = x_{(1)} - b_{n-1}^*/(nc-1) \tag{37}$$

$$b_k^* = (1/U_{n-1}) \{(c+1)\sum_{i=1}^{n-2} B_i \; x_{(i)} + (2c-1) \; B_{n-1} \; x_{(n-1)}$$

$$- [(nc-1)/(nc)] \; (nc-2-U_{n-1}) \; x_{(1)}\} \tag{38}$$

where

$$U_k = U_{n-1} = \frac{(nc-2)(nc-c-2) - nc(c-2) \; B_{n-1}}{(nc-1)(c+2)} \tag{39}$$

Summary of BLUEs. For shape parameter $c = .5$, 1, or 2, this thesis uses equations (34) and (35) to calculate the BLUEs for location parameter a and scale parameter b; however, the case $c = .5$ and $n = 5$ is omitted, since then the BLUEs cannot be found. When $c = 1.5$, the BLUEs are given by equations (37) to (39). For $c = 2.5$, 3, 3.5, or 4, equations (17), (18), (21), (22) and (29) are used to calculate the BLUEs for a and b. Once the BLUEs have been computed, the K-S, A-D and C-VM test statistics can be modified to accomodate unspecified location and scale parameters. An example will help to illustrate the calculations involved.

Example 1. In Table I the data listed under the x_i column was generated from a Pareto distribution of shape parameter $c = 2.5$, using equation (47) in the next chapter. Suppose it is desired to find the BLUE estimators $\hat{a}$ and $\hat{b}$ based on this particular random sample of size $n = 10$. Since in this case it is known that $c = 2.5$, the BLUEs will be computed from equations (17) and (18). One procedure to accomplish this is as follows:

Step 1. Arrange the x_i sample values in order from smallest to largest. The resulting order statistics (20:70) are listed under the $x_{(i)}$ column of Table I.

Table I

CALCULATION OF BLUES

i	x_i	$x_{(i)}$	C_i	B_{i-1}	B_i	$B_i x_{(i)}$
1	1.7986	1.0095	.9200	1.0000	.9200	.9287
2	1.0684	1.0586	.9111	.9200	.8382	.8873
3	1.3725	1.0684	.9000	.8382	.7544	.8060
4	1.1779	1.1267	.8857	.7544	.6682	.7529
5	1.4743	1.1779	.8667	.6682	.5791	.6821
6	1.0095	1.3725	.8400	.5791	.4864	.6676
7	4.8304	1.4743	.8000	.4864	.3891	.5737
8	1.0586	1.7986	.7333	.3891	.2854	.5133
9	1.1267	3.9974	.6000	.2854	.1712	.6844
10	3.9974	4.8304	.2000	.1712	.0342	.1652

$$D = (c+1)\sum_{i=1}^{n-1} B_i + (c-1)B_n = 17.8733$$

$$Y = (c+1)\sum_{i=1}^{n-1} B_i\, x_{(i)} + (c-1)B_n\, x_{(n)} - Dx_{(1)} = 4.9407$$

$$\hat{a} = x_{(1)} - Y/[(nc-1)(nc-2)-ncD] = .9625$$

$$\hat{b} = (x_{(1)}-\hat{a})(nc-1) = 1.128$$

Step 2. Compute each B_i for $i=1,2,\cdots,n$ using equation (29). Thus:

For $i=1$, $B_1 = [1-2/2.5(10-1+1)]B_0 = (1-2/25.0)(1.000) = .9200$

For $i=2$, $B_2 = [1-2/2.5(10-2+1)]B_1 = (1-2/22.5)(.9200) = .8382$

.

.

.

For $i=10$, $B_{10} = [1-2/2.5(10-10+1)]B_9 = (1-2/2.5)(.1712) = .0342$

Table I lists all of the values of $C_i = 1 - 2/c(n-i+1)$ and $B_i = C_i B_{i-1}$ as computed from equation (29).

Step 3. Use the B_i to compute D from equation (21):

$$D = (c+1)(B_1+B_2+\cdots+B_9) + (c-1)B_{10}$$

$$= (2.5 + 1)(.9200+ .8382+\cdots+ .1712) + (2.5 - 1)(.0342)$$

$$= (3.5)(5.092) + (1.5)(.0342) = 17.8733$$

Step 4. Use the $x_{(i)}$, D, and B_i values to compute Y from equation (22). Table I lists the values of $B_i x_{(i)}$:

$$Y = (c+1)[B_1x_{(1)}+ B_2x_{(2)}+ \cdots+ B_9x_{(9)}] + (c-1)B_{10}x_{(10)} -Dx_{(1)}$$

$$= (3.5)(.9287+ .8873+ \cdots+ .6844)$$

$$+ (1.5)(.1652) - 17.8733(1.0095)$$

$$= (3.5)(6.496) + .2478 - 18.0431 = 4.9407$$

Step 5. Use Y and D to compute $\hat{a}$ from equation (17):

$$\hat{a} = x_{(1)} - Y/[(nc-1)(nc-2)-ncD]$$

$$= 1.0095 - (4.9407)/[(25-1)(25-2) -25(17.8733)]$$

$$= 1.0095 - 4.9407/105.1675 = .9625$$

Step 6. Use $\hat{a}$ to compute $\hat{b}$ from equation (18):

$$\hat{b} = (x_{(1)}-\hat{a})(nc-1) = (1.0095 - .9625)(25 - 1) = 1.128$$

In this example, then, the BLUEs for a and b are $\hat{a} = .9625$ and $\hat{b} = 1.128$. (The x_i values were actually generated from a Pareto distribution with a = b = 1 and c = 2.5). Once the BLUEs have been computed, the test statistics can be appropriately modified.

Modified Test Statistics

At the end of Chapter II, the standard forms of the Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Cramer-von Mises (C-VM) test statistics were presented. To use these "unmodified" statistics with their existing critical value tables, all parameters must be specified. When unknown location and scale parameters are involved, the test statistics must be modified to generate new critical value tables before they will produce accurate results. This section shows how to calculate the modified test statistics using an ordered sample and the BLUEs described in the preceding section. The notation and approach are adapted from Littell, McClave, and Offen (36:259-260).

Hypothesized Pareto CDF. Before computing the modified test statistics, the hypothesized Pareto CDF must be calculated for each value of the random sample. Let $x_1, x_2, \cdots, x_n$ be a random sample from the Pareto distribution with unknown location and scale parameters a and b, and known shape c; and let $x_{(i)}$ denote the ith order statistic (20:70). The

appropriate BLUEs for location a and scale b (computed from the previous section), the specified shape c, and the n ordered Pareto deviates, $x_{(i)}$, are substituted into equation (15) to calculate the hypothesized Pareto CDF:

$$P_i = F(x_{(i)};\hat{a},\hat{b},c) = 1 - [1 + (x_{(i)}-\hat{a})/\hat{b}]^{-c} \qquad (40)$$

for i = 1,2, ··· ,n. Note that for a given shape c (e.g., c=2.5 or c=4) and sample size n (e.g., n=10 or n=30), a specific, fixed pair of location and scale values (e.g., a=b=1 or a=0, b=1) is used to produce the random Pareto deviates needed to compute the hypothesized CDF. This can be done without loss of generality because, as discussed in Chapter II, the use of invariant estimators (in this case the BLUEs) for location and scale ensures that the distribution of the test statistic depends only on the shape c and sample size n, and is independent of location and scale (36:260).

<u>Example 2</u>. In Example 1, the BLUEs for location a and scale b were found from a sample of size n=10 generated from a Pareto distribution having shape c=2.5. In this example, the same sample of values $x_1,x_2,\cdots,x_{10}$ will be used to compute the hypothesized Pareto CDF from equation (40). Table II contains the values obtained while making the calculations. The columns for x_i and $x_{(i)}$ are duplicated from Table I. The BLUEs $\hat{a}$ and $\hat{b}$ are as derived in Example 1.

Table II

CALCULATION OF HYPOTHESIZED PARETO CDF

i	x_i	$x_{(i)}$	M_i	N_i	O_i	P_i
1	1.7986	1.0095	.0470	.0417	.9030	.0970
2	1.0684	1.0586	.0961	.0852	.8151	.1849
3	1.3725	1.0684	.1059	.0939	.7990	.2010
4	1.1779	1.1267	.1642	.1456	.7119	.2881
5	1.4743	1.1779	.2154	.1910	.6460	.3540
6	1.0095	1.3725	.4100	.3635	.4607	.5393
7	4.8304	1.4743	.5118	.4537	.3925	.6075
8	1.0586	1.7986	.8361	.7412	.2500	.7500
9	1.1267	3.9974	3.0349	2.6905	.0382	.9618
10	3.9974	4.8304	3.8679	3.4290	.0242	.9758

$M_i = x_{(i)} - \hat{a} = x_{(i)} - .9625$

$N_i = M_i \,/\, \hat{b} = M_i \,/\, 1.128$

$O_i = (1 + N_i)^{-c} = (1 + N_i)^{-2.5}$

Hypothesized Pareto CDF: $P_i = 1 - O_i$

Modified K-S Statistic. After computing all n of the values of P_i from equation (40), the modified Kolmogorov-Smirnov test statistic is found from equation (4) by substituting P_i in place of z_i in equation (5). Thus the modified test statistic in computational form is:

$$D = \max\ (D^+, D^-) \tag{41}$$

where

$$D^+ = \sup_{1\le i\le n}\ [(i/n)-P_i] \quad \text{and} \quad D^- = \sup_{1\le i\le n}\ [P_i-(i-1)/n] \tag{42}$$

Table III

CALCULATION OF MODIFIED K-S TEST STATISTIC

i	$x_{(i)}$	P_i	i/n	$(i-1)/n$	D_i^+	D_i^-
1	1.0095	.0970	.1	.0	.0030	.0970
2	1.0586	.1849	.2	.1	.0151	.0849
3	1.0684	.2010	.3	.2	.0990	.0010
4	1.1267	.2881	.4	.3	.1119	−.0119
5	1.1779	.3540	.5	.4	(.1460)	−.0460
6	1.3725	.5393	.6	.5	.0607	.0393
7	1.4743	.6075	.7	.6	.0925	.0075
8	1.7986	.7500	.8	.7	.0500	.0500
9	3.9974	.9618	.9	.8	−.0618	(.1618)
10	4.8304	.9758	1.0	.9	.0242	.0758

$$D_i^+ = (i/n) - P_i = i/10 - P_i$$

$$D^+ = \sup\,[(i/n)-P_i] = .1460$$

$$D_i^- = P_i - (i-1)/n = P_i - (i-1)/10$$

$$D^- = \sup\,[P_i - (i-1)/n] = .1618$$

K-S Statistic: $D = \max\,(D^+, D^-) = .1618$

<u>Example 3</u>. Once the hypothesized Pareto CDF is computed, the values can be used to calculate the modified K-S test statistic. Table III continues the previous examples by showing the computations involved in calculating the modified K-S test statistic. As before, the calculations are based on the n=10 order statistics introduced in example 1, and the values P_i of the hypothesized Pareto CDF as computed in example 2.

Table IV

CALCULATION OF MODIFIED A-D TEST STATISTIC

j	P_j	P_{n+1-j}	L_j	M_j	N_j	$(2j-1)N_j$
1	.0970	.9758	-2.3330	-3.7214	-6.0544	-6.0544
2	.1849	.9618	-1.6879	-3.2649	-4.9528	-14.8584
3	.2010	.7500	-1.6045	-1.3863	-2.9908	-14.9540
4	.2881	.6075	-1.2444	-.9352	-2.1796	-15.2572
5	.3540	.5393	-1.0385	-.7750	-1.8135	-16.3215
6	.5393	.3540	-.6175	-.4370	-1.0545	-11.5995
7	.6075	.2881	-.4984	-.3398	-.8382	-10.8966
8	.7500	.2010	-.2877	-.2244	-.5121	-7.6815
9	.9618	.1849	-.0389	-.2040	-.2429	-4.1293
10	.9758	.0970	-.0245	-.1020	-.1265	-2.4035

$$\sum_{j=1}^{n} (2j-1)N_j = -104.1559$$

$L_j = \ln P_j$ $\qquad$ $M_j = \ln(1-P_{n+1-j})$ $\qquad$ $N_j = L_j + M_j$

$$A^2 = -n - (1/n) \sum_{j=1}^{n} (2j-1)[\ln P_j + \ln(1-P_{n+1-j})]$$

$$= -10 - (1/10)(-104.1559) = .4156$$

<u>Modified A-D Statistic</u>. The modified Anderson-Darling test statistic is computed by substituting P_j from equation (40) in place of z_j in equation (9). Thus the computational form of the modified A-D test statistic is:

$$A^2 = -n - (1/n) \sum_{j=1}^{n} (2j-1)[\ln P_j + \ln(1-P_{n+1-j})] \qquad (43)$$

<u>Example 4</u>. Table IV shows the calculations involved in finding the value of the modified A-D test statistic. The P_j values are as computed in example 2.

Table V

CALCULATION OF MODIFIED C-VM TEST STATISTIC

j	F_j	$\frac{2j-1}{2n}$	$P_j - \frac{(2j-1)}{2n}$	$[P_j - \frac{(2j-1)}{2n}]^2$
1	.0970	.05	.0470	.0022
2	.1849	.15	.0349	.0012
3	.2010	.25	-.0490	.0024
4	.2881	.35	-.0619	.0038
5	.3540	.45	-.0960	.0092
6	.5393	.55	-.0107	.0001
7	.6075	.65	-.0425	.0018
8	.7500	.75	.0000	.0000
9	.9618	.85	.1118	.0125
10	.9758	.95	.0258	.0007
			$\sum_{j=1}^{n}$ =	.0339

$$W^2 = [1/(12n)] + \sum_{j=1}^{n}[P_j - (2j-1)/2n]^2$$

$$= (1/120) + .0339 = .0423$$

Modified C-VM Statistic. The computational form of the modified Cramer-von Mises test statistic is found from equation (7) by substituting P_j for z_j:

$$W^2 = [1/(12n)] + \sum_{j=1}^{n}[P_j - (2j-1)/2n]^2 \quad (44)$$

Example 5. Table V shows the calculations involved in finding the value of the modified C-VM test statistic. The P_j values are as computed in example 2.

Chapter Summary

Several applications for the Pareto distribution have been found in economics and operations research. It has played a major role in investigating the distributions of city population size, natural resources, stock price fluctuations, and oil field locations. Other studies show the Pareto can be used to model phenomena which may apply to Air Force interests, such as time-to-failure of equipment components, maintenance service times, nuclear fallout dispersion, and error clusters in communications circuits.

There are three basic forms of the Pareto distribution, each of which is a special case of the three-parameter form. The greater generality of the three-parameter form allows the Pareto distribution to be more useful in practical application. Various methods have been explored for estimation of Pareto parameters; but the best linear unbiased estimator (BLUE) is the only estimator known to possess the required invariance property for the three-parameter form.

For shape parameter $c = .5$, 1, or 2, the BLUEs are computed from equations (34) and (35). When $c = 1.5$, the BLUEs are given by equations (37) to (39). For $c = 2.5$, 3, 3.5, or 4, the BLUEs are computed from equations (17), (18), (21), (22), and (29). The BLUEs are used to compute the hypothesized distribution function from equation (40). The modified K-S, A-D, and C-VM test statistics can then be found using the methods presented in the next chapter.

IV. METHODOLOGY

Chapter Overview

This chapter describes the basic principles and specific procedures used to satisfy the research objectives of this thesis. Foremost is the Monte Carlo method used to generate the critical value tables of the modified K-S, A-D, and C-VM goodness-of-fit tests for the three-parameter Pareto distribution when only the shape parameter is specified.

Basic Principles

This section deals with some of the basic principles used to generate critical values. It begins with an overview of the Monte Carlo method in general. Next is discussed the inverse transform technique used to generate random Pareto deviates. Then the selection of critical values is discussed. Finally, the use of plotting positions to determine percentiles is explained.

The Monte Carlo Method. Mathematics can be divided into theoretical and experimental categories. The primary distinction is that "theoreticians deduce conclusions from postulates, whereas experimentalists infer conclusions from observations" (21:1). The Monte Carlo method is a branch of experimental mathematics involving experiments using random

numbers. It has been used extensively in statistical analysis, operational research, nuclear physics, and several other fields where there are problems not easily solved by theoretical mathematics alone (21:2).

An important feature of the Monte Carlo method is its usual reliance on computers to simulate random processes (10:2). Also known as the method of statistical trials, it is basically a system of techniques which allows the modeling of random processes conveniently by digital computer. Before the advent of the computer, a study of a random process was considered to be complete when it was reduced to an analytical description. The computer has now made it convenient in many cases to solve an analytical problem by reducing it to a random process and then simulating that process (10:vii). Thus a basic principle of the method involves simulating statistical experiments through computational techniques, and then analysing numerical characteristics observed from these experiments (10:ix). For this reason, the Monte Carlo method can be defined as "the construction of an artificial random process possessing all the necessary properties, but which is in principle realizable by means of ordinary computational apparatus" (10:2).

The Monte Carlo method is typically used to solve problems of two basic types. A deterministic problem has no direct association with random processes. In this case the Monte Carlo method is often used when the problem can be

formulated in theoretical language but cannot be solved by theoretical means. Usually the approach is to recognize the underlying problem structure as resembling some apparently unrelated random process, and then solve the deterministic problem numerically by an appropriate Monte Carlo simulation.

In the case of a probabilistic problem, the Monte Carlo method is directly concerned with the behavior and outcome of random processes. The approach is to observe random variates, chosen so that they directly simulate the physical random processes of the original problem. The desired solution is then inferred from the behavior of the random numbers (21:2-4). The latter Monte Carlo approach was used in this thesis to generate the critical value tables for the goodness-of-fit tests.

The main weakness in the Monte Carlo method is that the answers it produces are to some degree uncertain since they are inferred from raw observational data consisting of random numbers. This weakness must be accounted for because:

> Whenever one is inferring general laws on the basis of particular observations associated with them, the conclusions are uncertain inasmuch as the particular observations are only a more or less representative sample from the totality of all observations which might have been made. Good experimentation tries to ensure that the sample shall be more rather than less representative . . . [Monte Carlo answers] can nevertheless serve a useful purpose if we can manage to make the uncertainty fairly negligible, that is to say to make it unlikely that the answers are wrong by very much [21:4-5].

Thus there is usually no cause for concern if the uncertainty is negligible for practical purposes.

One way of reducing uncertainty is to base the Monte Carlo analysis on a larger number of observations. However, economic and time constraints must be considered. "Broadly speaking, there is a square law relationship between the error in an answer and the requisite number of observations; to reduce it tenfold calls for a hundredfold increase in the observations, and so on" (21:5). Therefore, to avoid using an inordinate amount of computer time, and to conserve financial resources, this thesis follows the common practice (9;43;49;52;54) of using 5000 repetitions rather than, say, 10000 in performing the Monte Carlo analysis.

The Inverse Transform Technique. To apply the Monte Carlo method to the problem at hand requires random samples from the Pareto distribution. The most practical way to obtain such samples is to use a computer program to produce a group of n numbers that seem to come from a Pareto population. In terminology adapted from Conover (13:323-324,360), these n numbers are called "random Pareto deviates" because they are deliberately generated to resemble observations on independent Pareto random variables. Previous AFIT theses (9;43;49; etc.) involved distributions for which computer programs to generate random samples were already available from the International Mathematical Statistics Library

(IMSL). IMSL does not contain a similar subroutine for the Pareto distribution; therefore, a computer program needed to be written to generate random Pareto deviates.

One common method of using a computer to generate random samples from a given distribution is to first generate a uniform random sample on (0,1) and then transform it into a new sample having the desired distribution. This method, called the inverse transform technique, uses the fact that the random variable $R = F(X)$ is uniformly distributed on (0,1), where X is a random variate (5:293-298). Thus, every variate is related to the uniform variate on (0,1) through its own inverse distribution function (26:22). Therefore, a set of uniformly distributed random numbers is required to generate a random sample from the Pareto distribution.

Conveniently, most random number generators are designed to generate random numbers which are uniformly distributed on the interval (0,1) (5:293). Hence, the inverse transform technique can be directly applied to a set of these random numbers to generate random Pareto deviates. However, the technique requires that for each random number r, the equation $r = F(x)$ must be solved for the corresponding value of $x = F^{-1}(r)$. Therefore the technique is practical only when the CDF F(x) has an inverse which can be computed explicitly (5:294). Fortunately, the inverse transformation for the Pareto distribution can easily be expressed in closed form.

The inverse transform technique can be accomplished by the following four-step procedure (5:294-295):

Step 1. Compute the cumulative distribution function (CDF) of the desired random variable X. In this case, the CDF is the three-parameter Pareto CDF, given by equation (15) and repeated here for convenience:

$$F(x) = 1 - [1 + (x-a)/b]^{-c} \quad \text{for } x \geq a;\ b,c > 0$$

Step 2. Set $F(X) = R$ on the range of X, where X represents a random Pareto variable. This then becomes:

$$1 - [1 + (X-a)/b]^{-c} = R \quad \text{for } x \geq a \tag{45}$$

Since X is a random variable (with the Pareto distribution in this case), then R is also a random variable. In fact, R has a uniform distribution over the interval (0,1) (5:295).

Step 3. Solve F(X) in terms of R to find $X = F^{-1}(R)$. In this case the inverse is found by solving equation (45):

$$1 - [1 + (X-a)/b]^{-c} = R$$

$$[1 + (X-a)/b]^{-c} = 1 - R$$

$$[b/b + (X-a)/b]^{-c} = 1 - R$$

$$(b + X - a)/b = (1 - R)^{-1/c}$$

$$b + X - a = b(1 - R)^{-1/c}$$

Therefore $$X = (a - b) + b(1 - R)^{-1/c} = F^{-1}(R) \tag{46}$$

Equation (46) is called a "random variate generator" (5:295) for the Pareto distribution. As explained in the discussion following equation (40), a specific, fixed pair of location and scale values can be used to generate the required deviates without loss of generality. For this thesis, the Pareto deviates were generated using location and scale parameters of 1. Substituting a=b=1 into equation (46) gives:

$$\begin{aligned} X &= a - b + b(1 - R)^{-1/c} \\ &= 1 - 1 + 1(1 - R)^{-1/c} \\ &= (1 - R)^{-1/c} \qquad (47) \end{aligned}$$

Since R is uniformly distributed from 0 to 1, then so is 1-R; thus R can replace 1-R in equation (47) to yield the particular random variate generator used to produce the random Pareto variates for this thesis:

$$X = R^{-1/c} = (1/R)^{1/c} \qquad (48)$$

Step 4. Generate n uniform random numbers $R_1, R_2, \cdots, R_n$ and compute the n random Pareto deviates from equation (48). The random numbers used for this thesis were generated on the AFIT VAX/VMS computer system using the IMSL subroutine GGUBS. Like most random number generators (5:293), GGUBS is designed to generate random numbers which

are uniformly distributed on the interval (0,1). Therefore, the inverse transform technique was applied to these random numbers to generate random Pareto deviates.

In step 3 of the inverse transform procedure, the choice of the location and scale values is arbitrary, and 1 was used here for convenience. It should be noted, however, that the deviates can be easily transformed into deviates from a different Pareto distribution (i.e., one having the same shape c but different location a' or scale b'). The transformation stems from the fact that all variates having the same shape can be expressed in terms of the variate having location 0 and scale 1, as follows (26:21-22):

$$X_{a,b} = b\, X_{0,1} + a \tag{49}$$

where $X_{a,b}$ denotes a Pareto variate with location a and scale b and $X_{0,1}$ is a Pareto variate with location 0 and scale 1. The transformation to the different variate is then found by expressing the given variate in terms of the 0,1 variate, since:

$$X_{a,b} = b\, X_{0,1} + a \quad \text{implies} \quad X_{0,1} = (X_{a,b} - a)/b$$

Thus $X_{a',b'} = b'\, X_{0,1} + a' = b'[(X_{a,b} - a)/b] + a'$ (50)

Therefore, given a variate having a specific pair of

values for location and scale, equation (50) can be used to transform the variate to one having a different pair of location and scale parameters. For example, the transformation from a variate having location and scale a=b=1 to one having location a'=2 and scale b'=3 is given by:

$$X_{2,3} = 3X_{0,1} + 2 = 3[(X_{1,1} - 1)/1] + 2 = 3X_{1,1} - 1$$

The random Pareto deviates generated by the inverse transform technique were used ultimately to compute values of the modified K-S, A-D, and C-VM test statistics. However, these test statistics can only be useful if their distribution functions are at least partially known (13:31). Thus, many test statistics were computed to determine the empirical distribution. Critical values were then identified using a plotting positions technique. Before examining the plotting positions technique, it may be helpful to understand how critical values are chosen.

Identifying Critical Values. The use of random deviates to generate critical value tables is based on the concept of hypothesis testing mentioned in Chapter II. Each group of n Pareto deviates represents a simulated sample from a parameter-specified Pareto distribution. This makes the null hypothesis "H_0: H(x) = the Pareto CDF" true for each sample of n random Pareto deviates. For each of the three

tests (K-S, A-D, and C-VM), equations (41) - (44) were used to compute 5000 independent values of the test statistic under the condition that H_0 is true (13:361). These 5000 values were then arranged in ascending order to form sets of 5000 order statistics. To determine critical values from these 5000 statistics (15000 total for all three tests), it is necessary to identify somehow the "critical region", i.e., the set of all values of the test statistic that would result in the erroneous decision to reject the true null hypothesis (13:78). Once the critical region is identified, then the critical values can be selected according to a desired "level of significance", or α, which is the maximum probability of rejecting a true null hypothesis. Since the use of random Pareto deviates to compute the test statistics ensures that H_0 is true, α can be found by determining the probability that the test statistic will assume a value that falls within the critical region (13:78).

Since H_0 is true and α is the maximum probability of rejecting H_0, then the *minimum* probability of correctly accepting H_0 is $1-\alpha$. This value of $1-\alpha$ represents a certain percentile of the 5000 ordered test statistic values. For example, the 99th percentile is some number that the test statistic will exceed with probability .01 or less and will be less than with probability .99 or less (13:29). It is this percentile relationship that is used to select critical values from the 5000 test statistics.

One possible method of using the percentiles to determine critical values is to simply select the test statistic value corresponding to the desired percentile level and make that the critical value. For example, under this method, out of a set of 5000 ordered test statistic values, the critical value for the 90th percentile would simply be the 4500th value (52:6). This method has some disadvantages, however, especially when the test statistics, which represent a discrete distribution, are used to determine critical values for a continuous distribution. More recently, the plotting position technique has become popular as a more accurate method of selecting critical values for continuous distributions (43:7).

The Plotting Positions Technique. The plotting positions technique is one popular method of determining percentiles of the distribution underlying a se of n ordered sample values (24:1619; 25:317). The technique involves using a large number of discrete values of the ordered test statistics and locating them on a continuous spectrum by representing the spaces between them as piecewise linear functions. This makes it possible to linearly interpolate the desired percentiles between discrete values of the test statistics, thus obtaining more accurate critical values (43:7; 52:6).

Each ordered value may be assigned a plotting position

which is its cumulative probability, thus allowing each order statistic to be mapped onto a probability scale from 0 to 1. As seen from equation (2), the distribution function of these n observations is a step function which jumps from $(i-1)/n$ to i/n at the ith order statistic of the sample. However, if the plotting position i/n is used, the largest value cannot be plotted, while if $(i-1)/n$ is used, the smallest value cannot be plotted (24:1615). Therefore, numerous alternative plotting conventions have been proposed, most of which have been summarized by Harter (24), who presents various arguments for and against each. Harter also conducted a Monte Carlo analysis of plotting positions for several distributions and concluded that ". . . the optimum choice of plotting positions depends not only on the purpose of the investigation, but also (definitely) on the distribution of the variable under consideration" (25:342).

While Harter made no specific recommendation for the Pareto, he did observe that, "As samples increase above a sample size of 20, the differences among the positions determined by any method of estimation decrease to the point where they are practically unimportant" (24:1621). He also noted that "in practice, plotting positions differ little compared with the randomness of the data" (24:1622). Since this thesis employed 5000 independent values of each test statistic, well in excess of the 20 cited by Harter, use of a single plotting convention seems justified.

The plotting convention selected for this thesis is the median rank, which is closely approximated by the plotting position (24:1617):

$$Y_i = (i-0.3)/(n+0.4) \qquad (51)$$

where $i = 1,\cdots,n$ and for this thesis, n=5000. Thus each Y_i value lies in the interval (0,1). The median ranks position yields median unbiased estimates of x_i for a specified $F(x_i)$ and of $F(x_i)$ for a specified x_i (24:1625). Also, in highly skewed distributions, the median ranks position tends to be more accurate than other conventions (31:300). Another advantage is that values of the median ranks have been tabulated for sample sizes of 1 to 50, i.e., n = 1(1)50 (31:486-489).

A detailed illustration showing how to use plotting positions to determine critical values was presented by Ream (43:11-23), and will only be summarized here. In graphical terms, the technique effectively plots the 5000 ordered test statistic values $X_{(1)},X_{(2)},\cdots,X_{(5000)}$ along the abscissa (horizontal) axis and the 5000 plotting position values $Y_1,Y_2,\ldots,Y_{5000}$ computed from equation (51) along the ordinate (vertical) axis. These values are assigned to positions 2 to 5001 on their respective axes. On the vertical axis, the interval [0,1] is completed by entering the endpoints $Y_0 = 0$ at the 1st position and $Y_{5001} = 1$ at the

5002nd position. The corresponding endpoints on the horizontal axis are found by linear extrapolation. Thus, in using the computer to program this technique, the arrays corresponding to the horizontal and vertical axes are each composed of 5002 entries, i.e., the original 5000 values and two extrapolated endpoints.

To map the collection of 5000 discrete values onto a fully continuous line between 0 and 1 requires extrapolation of the endpoints of the plotting axes. The first point on the horizontal axis, $X_{(0)}$, is computed by linearly extrapolating from the second and third points (i.e., the first and second order statistics), subject to a non-negativity restriction. Extrapolation is performed by using the standard linear slope-intercept formula $Y = mX + b$ to compute the endpoints $X_{(0)}$ and $X_{(5001)}$. To find the first endpoint on the horizontal axis, the slope is calculated by:

$$m = \frac{Y_2 - Y_1}{X_{(2)} - X_{(1)}} \tag{52}$$

and the intercept is:

$$b = Y_1 - m X_{(1)} \tag{53}$$

Then the lower endpoint $X_{(0)}$ is found by:

$$X_{(0)} = (Y_0 - b)/m = (0-b)/m = -b/m$$

The nonnegativity restriction means that whenever $-b/m < 0$, then $X_{(0)}$ is simply set to 0. Thus:

$$X_{(0)} = \max(0, -b/m) \tag{54}$$

The higher endpoint $X_{(5001)}$ is found in the same way as the lower endpoint. The slope is

$$m = \frac{Y_{5000} - Y_{4999}}{X_{(5000)} - X_{(4999)}} \tag{55}$$

and the intercept is:

$$b = Y_{4999} - m\,X_{(4999)} \tag{56}$$

Then the second endpoint $X_{(5001)}$ is extrapolated by:

$$X_{(5001)} = (Y_{5001} - b)/m = (1-b)/m \tag{57}$$

Once the endpoints are added to the abscissa and ordinate axes, the 5002 discrete points on the graph are "connected" by straight lines, thus producing a completely continuous, piecewise linear function. The range of this continuous function is the interval [0,1] and contains the 5000 median rank values as well as the endpoints 0 and 1.

Its domain contains the set of 5000 test statistic values and their 2 extrapolated endpoints.

As shown in Figure 5, the desired critical value for a given percentile is found by linearly interpolating between two of the 5002 points used to construct the now continuous graph. For example, to find the 95th percentile (α = .05), the largest plotting position Y_j is found such that $Y_j \leq .95$; thus Y_{j+1} is the first position greater than .95. Then the critical value corresponding to the 95th percentile is found by linearly interpolating between the points $(X_{(j)}, Y_j)$ and $(X_{(j+1)}, Y_{j+1})$ using the formulas:

$$m = \frac{Y_{j+1} - Y_j}{X_{(j+1)} - X_{(j)}} \tag{58}$$

$$b = Y_j - m X_{(j)} \tag{59}$$

$$C_p = (p - b)/m \tag{60}$$

where C_p is the critical value for the the 100pth percentile. For this thesis, critical values were calculated for p = .80, .85, .90, .95, and .99, corresponding to the levels of significance α = .20, .15, .10, .05, and .01.

The specific plotting position procedure performed for this thesis is described in step 7 of the next section.

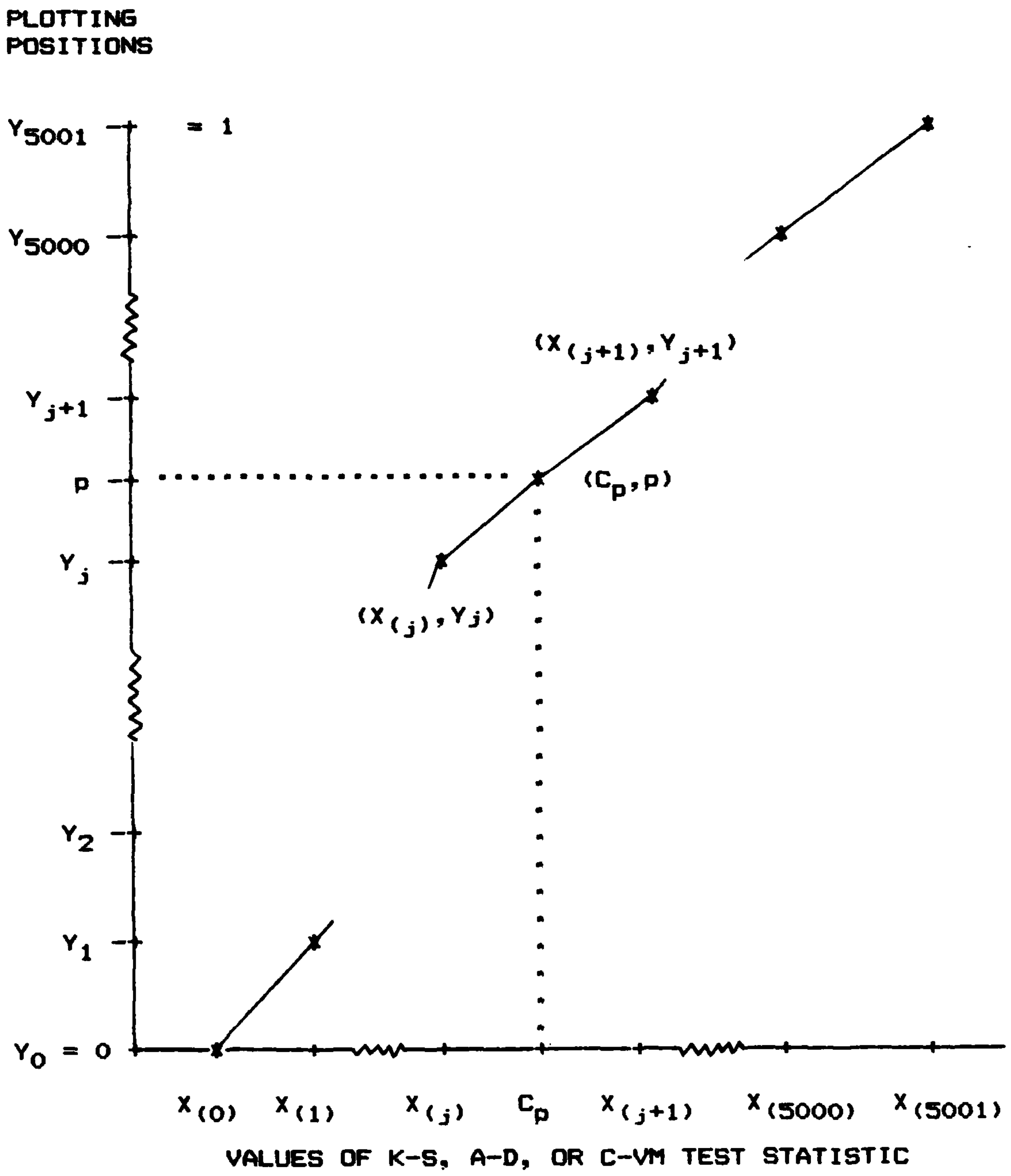


Fig 5. Using Test Statistics $X_{(j)}$ and Plotting Positions Y_j to Find Critical Value C_p for the 100(p)th Percentile (p = .99, .95, .90, .85, .80).

Specific Procedures

By applying the basic principles and techniques described in the previous section, the K-S, A-D, and C-VM tests were modified to produce new goodness-of-fit tests for the Pareto distribution.

The research effort was performed in three stages, each corresponding to one of the three research objectives listed in Chapter I. The first stage consisted of a nine-step Monte Carlo simulation procedure to produce critical value tables for the modified K-S, A-D, and C-VM tests. The second stage of the research compared the powers of the three modified tests using eight alternative distributions. Finally, a regression analysis was performed to determine the functional relationship between the critical values and the shape parameters. Computer programs were written to accomplish the first two stages. The third stage was performed manually by using a hand calculator to compute linear relationships by the method of least squares.

Stage 1: Generating Critical Value Tables. During the first stage, critical value tables were generated using Monte Carlo simulation. A FORTRAN computer program was written for this purpose and is contained in Appendix A. The accompanying flow chart illustrates the logic flow of the program. The following nine steps outline the procedure used:

Step 1 - Generate the Data. Random deviates for a given sample size n were generated from a specified Pareto distribution by using the IMSL routine GGUBS to generate n random numbers, and then applying the inverse transform technique (equation 48).

Step 2 - Order the Data. Next, the n random deviates $x_1, x_2, \cdots, x_n$ were converted to order statistics $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ by arranging them in ascending order using the IMSL subroutine VSRTA.

Step 3 - Estimate the Parameters. The ordered Pareto deviates were then used to find the best linear unbiased estimates of the scale and location parameters as explained in the "Summary of BLUEs" section of Chapter III.

Step 4 - Compute the Hypothesized CDF. The estimated parameters found in step 3 were used with the n ordered Pareto deviates from step 2 to calculate the hypothesized cumulative distribution function (CDF) P_i for $i=1,2,\cdots,n$ (equation 40 in chapter III).

Step 5 - Calculate the Test Statistics. Based on the hypothesized CDF and the BLUEs, the modified K-S, A-D, and C-VM statistics were next calculated using equations (42), (43), and (44).

Step 6 - Generate 5000 Statistics. Each of these five steps were repeated 5000 times to generate 5000 independent K-S, A-D, and C-VM statistical values $X_1, X_2, \cdots, X_{5000}$.

Step 7 - Find the Critical Values. For each of the three tests, the 5000 statistics were ordered as in step 2. Using the median ranks plotting position technique (equation 51), the 80th, 85th, 90th, 95th, and 99th percentiles of the distributions of each test statistic were calculated by linear interpolation. These percentiles correspond, respectively, to the .20, .15, .10, .05, and .01 levels of significance and served as the critical values for the modified K-S, A-D, and C-VM goodness-of-fit tests. The specific step-by-step process was to:

a. Use the IMSL subroutine VSRTA to order the 5000 test statistics, thus forming the 5000 order statistics $X_{(1)}, X_{(2)}, \cdots, X_{(5000)}$.

b. Use equation (51) to compute the 5000 plotting positions $Y_1, Y_2, \cdots, Y_{5000}$. Also, set $Y_0 = 0$ and $Y_{5001} = 1$.

c. Use equations (52), (53), and (54) to find $X_{(0)}$. Similarly, use equations (55), (56), and (57) to find $X_{(5001)}$.

d. For a given p, find the largest Y_j such that $Y_j \leq p$; then use equations (58), (59), and (60) to find the critical value C_p representing the 100(p)th percentile. Repeat this step for p = .80, .85, .90, .95, and .99.

Step 8 - Repeat for Sample Sizes. To evaluate the effect of sample size on the critical values, steps 1 through 7 were repeated for each sample size n. This thesis

followed the common practice (9:15) of using sample sizes of n equal to 5, 10, 15, 20, 25, and 30.

Step 9 - Repeat for Shape Parameters. Steps 1 through 8 were repeated for specified shape parameters 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, and 4.0. The critical values were then arranged into tabular form and appear in Chapter V, Tables VI - VIII.

Stage 2: Comparing Power. The second stage of the research compared the powers of the modified K-S, A-D, and C-VM tests against the Chi-square to determine which test can best detect a false Pareto distribution hypothesis. As explained in Chapter II, the power of a statistical test is the probability of correctly rejecting a false null hypothesis. The null hypothesis that a set of sample deviates follows a Pareto distribution with a specified shape parameter was tested against the alternative hypothesis that the sample deviates follow some other distribution:

H_0: Sample deviates follow a Pareto CDF with shape c

H_1: They follow some other distribution

For this thesis, the power study was conducted for both c = 1 and c = 3.5 in the null hypothesis.

The Chi-square portion of the study was performed as described by Banks and Carson (5:352-356) using five

equiprobable (ie, $p = .20$) class intervals (or cells) with expected frequencies of 3 observations per cell for $n = 15$ and 5 per cell for $n = 25$. The endpoints of each cell were computed from the Pareto CDF (equation 15) as follows:

$$F(e_i) = 1 - [1 + (e_i - a)/b]^{-c} \qquad (61)$$

where e_1, e_2, e_3, e_4 represent the right endpoints (maximum value) of the first four cells. Since $F(e_i)$ is the cumulative area from 0 to e_i, then $F(e_i) = ip = .2i$, so equation (61) leads to:

$$.2i = 1 - [1 + (e_i - a)/b]^{-c}$$

$$[1 + (e_i - a)/b]^{-c} = 1 - .2i$$

$$1 + (e_i - a)/b = (1 - .2i)^{-1/c}$$

$$b + e_i - a = b(1 - .2i)^{-1/c}$$

$$e_i = a - b + b(1 - .2i)^{-1/c}$$

After substituting the BLUEs for location and scale into this last expression, the right endpoints were found by:

$$e_i = \hat{a} - \hat{b} + \hat{b}(1 - .2i)^{-1/c} \qquad (62)$$

Assuming a true Pareto null hypothesis, the four endpoints $e_1,\ldots,e_4$ essentially divide the real line into five equiprobable class intervals. Given a random sample, the

number of observations occuring within each cell were counted. The Chi-square test statistic was then computed by (5:350):

$$\chi^2 = \sum_{i=1}^{5} [(O_i - E)^2]/E \qquad (63)$$

where O_i is the number of observations occuring in cell i and $E = n/5$ is the expected frequency in each interval. The distribution of this test statistic approximately follows a chi-square CDF with s-1-k degrees of freedom (13:194) where s is the number of cells (i.e., s = 5) and k is the number of parameters estimated from the sample (i.e., k = 2).

Using the IMSL subroutines GGWIB, GGAMR, GGBTR, GGEXN, and GGNML, random deviates from different distributions of sample size n were generated. The alternate distributions used were, respectively, the Weibull at shape parameter 3.5, the Gamma at shape parameter 2.0, the Beta at parameters P = 2 and Q = 3, the exponential with mean = 2, and the normal distribution. Also tested were three sets of Pareto deviates generated by a FORTRAN subroutine. The first Pareto deviate set was generated using a = b = c = 1.0; the second set used a = 2, b = 3, and c = 3.5; the third used a = 10, b = 5, and c = 2.0. Five thousand random samples of size n were generated for each of the alternate distributions.

The K-S, A-D, C-VM, and Chi-square test statistics were then calculated under the null hypothesis that the random

deviates follow the Pareto distribution with specified shape $c = 1.0$ or 3.5. To determine whether to reject the null hypothesis, the calculated K-S, A-D, and C-VM statistics were compared to the corresponding critical value obtained in stage one. The computed Chi-square test statistic was compared against two sets of critical values. The first set was taken from a standard table of Chi-square critical values (13:432) based on 2 degrees of freedom. The second set of critical values was generated by using equations (62) and (63) and applying the 9-step, 5000-repetition Monte Carlo procedure described in the previous section.

This procedure of comparing test statistics against critical values was repeated 5000 times for each distribution and test. The number of times each statistic exceeded the respective critical value was counted for each sample size. This total, representing the number of rejections of the null hypothesis, was divided by the total number of tests performed (5000), to yield an hypothesis rejection quotient. For a random sample generated from the hypothesized Pareto distribution, the quotient represents the rate of erroneous rejection of a true null hypothesis; thus, it is expected to be approximately the level of significance α, which is the probability of committing a Type I error (13:78). In those cases involving random samples generated from an alternative distribution, the quotient represents the power of the test, since it approximates the probability of correctly rejecting

a false null hypothesis (13:79).

A FORTRAN program, written to compute the hypothesis rejection rates and accomplish the power study, is contained in Appendix B. Figure 7 in Appendix B shows how the program used the following 9-step process:

Step 1. Use IMSL or inverse transform to generate n random deviates from a selected distribution.

Step 2. Assume the null hypothesis that this set of n deviates follows the Pareto of given shape c = 1.0. Then perform steps 2-5 of the previous section to compute the values of the Chi-square (eqn 63) and modified K-S, A-D, and C-VM test statistics (eqns 42-44).

Step 3. For a given level of significance α, compare the test statistic value against the appropriate critical value found in the previous section. If the test statistic value equals or exceeds the critical value, H_0 is rejected.

Step 4. Repeat steps 1-3 5000 times, each time using a different seed to generate the deviates.

Step 5. Count the number of times H_0 was rejected and divide by 5000 to obtain the power.

Step 6. Repeat steps 1-5 for each alternative distribution considered.

Step 7. Repeat steps 1-6 for sample sizes n = 5, 15, and 25.

Step 8. Repeat steps 1-7 for α = .05 and .01.

Step 9. Repeat steps 1-8 using hypothesized Pareto shape c = 3.5. The power values were then arranged into tabular form and appear in Chapter V, Tables IX and X.

Stage 3: Determining Functional Relationship. The third and final stage of the research was to determine what (if any) functional relationship exists between the shape parameter and the critical values generated. This relationship can then be used to interpolate critical values corresponding to parameters not found in the generated tables.

To accomplish this stage, shape parameters and critical values were examined for linear relationships. In an attempt to "fit" the data to a line, a linear regression was performed using the method of least squares (13:263-271), which minimizes the sum of the squares of the deviations of the actual data points from the straight line of "best" fit (5:359-363). Where applicable, the correlation coefficient (13:250-251) was also found.

Linear regression is a capability available on many hand calculators currently on the market, so it was unnecessary to write a separate computer program to perform this function. For each level of significance and sample size, critical values from Tables VI - VIII were paired against a corresponding Pareto shape parameter. The

regression and correlation coefficients were then obtained manually by using the linear regression keys on a Texas Instruments TI-55-II calculator. The results are contained in Chapter V, Tables XI and XII.

Chapter Summary

The research for this thesis was performed by applying the Monte Carlo method using 5000 repetitions to generate critical value tables and a power study.

In stage 1, random Pareto deviates were generated by using the inverse transform technique, and 5000 test statistics were computed for each test. The median ranks plotting positions technique was then used to select critical values from the 5000 test statistics. In stage 2, the powers of the modified K-S, A-D, and C-VM tests were compared against the power of the Chi-square test. The calculations were performed by computer programs written to accomplish a 9-step Monte Carlo procedure. Stage 3 involved manual calculations based on the method of least squares to find linear relationships between shape parameters and critical values.

The results of this research are presented in the next chapter.

V. RESULTS AND APPLICATION

Chapter Overview

This chapter shows the results obtained from carrying out the methodology described in Chapter IV. In response to the three research objectives listed in Chapter I, tables of critical values for the modified K-S, A-D, and C-VM tests are presented. Also included are tables comparing powers of the K-S, A-D, and C-VM statistics against the Chi-square. Tables of regression coefficients are presented as well. The use of the tables is explained, and an example is described.

Critical Value Tables

Table VI contains critical values for the modified Kolmogorov-Smirnov Test. The modified Anderson-Darling critical values appear in Table VII. In Table VIII, the modified Cramer-von Mises critical values are presented. Critical values are presented for each level of significance α = .20, .15, .10, .05, and .01; sample sizes n = 5, 10, 15, 20, 25, and 30; and Pareto shape parameters .5, 1, 1.5, 2, 2.5, 3, 3.5, and 4. It is important to note that for shape c = 0.5, the presented critical values correspond to sample size n = 6 instead of n = 5. As explained in Chapter III, this exception is necessary since the BLUEs could not be computed for the case where c = .5, n = 5.

Table VI

CRITICAL VALUES FOR THE MODIFIED KOLMOGOROV-SMIRNOV TEST

		Pareto Shape Parameter c							
α	n	0.5*	1.0	1.5	2.0	2.5	3.0	3.5	4.0
.20	5*	.400	.318	.289	.286	.283	.286	.293	.297
	10	.255	.222	.217	.219	.222	.225	.228	.231
	15	.204	.184	.184	.185	.187	.191	.192	.197
	20	.175	.160	.160	.163	.167	.168	.170	.171
	25	.155	.144	.146	.148	.149	.153	.154	.155
	30	.142	.133	.135	.135	.138	.139	.142	.141
.15	5*	.426	.328	.296	.294	.293	.298	.306	.309
	10	.268	.230	.226	.228	.232	.236	.239	.242
	15	.214	.191	.191	.193	.196	.199	.203	.207
	20	.184	.167	.168	.172	.175	.176	.178	.179
	25	.163	.150	.152	.155	.157	.160	.161	.163
	30	.149	.138	.142	.142	.145	.146	.150	.149
.10	5*	.467	.341	.305	.306	.307	.314	.323	.327
	10	.284	.241	.236	.239	.245	.251	.253	.258
	15	.227	.200	.201	.204	.208	.212	.216	.219
	20	.196	.176	.176	.182	.185	.187	.188	.191
	25	.173	.158	.160	.164	.166	.171	.170	.173
	30	.159	.145	.149	.151	.153	.155	.161	.159
.05	5*	.525	.368	.321	.323	.328	.335	.349	.353
	10	.308	.257	.254	.258	.265	.272	.277	.282
	15	.248	.216	.217	.223	.227	.231	.238	.239
	20	.213	.188	.191	.197	.201	.206	.205	.209
	25	.189	.170	.174	.177	.180	.186	.189	.192
	30	.173	.156	.162	.165	.167	.169	.175	.174
.01	5*	.609	.407	.378	.363	.361	.369	.382	.391
	10	.348	.297	.290	.300	.308	.314	.322	.326
	15	.289	.247	.251	.258	.265	.266	.274	.282
	20	.247	.216	.221	.233	.233	.237	.238	.249
	25	.222	.201	.201	.208	.210	.220	.218	.225
	30	.204	.180	.187	.189	.196	.199	.207	.207

***NOTE:** For shape c = 0.5, critical values correspond to sample size n = 6 instead of n = 5.

Table VII

CRITICAL VALUES FOR THE MODIFIED ANDERSON-DARLING TEST

		Pareto Shape Parameter c							
α	n	0.5*	1.0	1.5	2.0	2.5	3.0	3.5	4.0
.20	5*	1.344	.736	.568	.546	.503	.494	.499	.497
	10	.780	.587	.544	.535	.541	.545	.540	.551
	15	.706	.589	.562	.559	.562	.568	.581	.588
	20	.684	.582	.571	.586	.591	.586	.599	.604
	25	.664	.588	.591	.585	.600	.608	.624	.621
	30	.674	.598	.607	.600	.606	.621	.638	.625
.15	5*	1.668	.835	.628	.602	.545	.532	.538	.537
	10	.875	.646	.594	.589	.588	.601	.597	.610
	15	.789	.645	.621	.612	.626	.630	.650	.659
	20	.764	.639	.629	.646	.656	.659	.661	.673
	25	.750	.653	.655	.652	.660	.672	.692	.694
	30	.756	.665	.679	.665	.678	.688	.708	.690
.10	5*	2.100	.966	.709	.671	.606	.585	.590	.599
	10	1.031	.726	.675	.655	.654	.678	.677	.691
	15	.917	.727	.705	.704	.705	.707	.748	.756
	20	.862	.732	.718	.734	.740	.747	.751	.755
	25	.853	.748	.742	.766	.750	.769	.788	.801
	30	.862	.756	.777	.768	.774	.776	.822	.791
.05	5*	2.903	1.237	.849	.791	.702	.683	.684	.687
	10	1.311	.886	.808	.783	.788	.805	.818	.835
	15	1.154	.891	.849	.853	.836	.852	.899	.927
	20	1.053	.874	.866	.898	.902	.917	.917	.925
	25	1.055	.915	.910	.940	.904	.926	.952	.987
	30	1.070	.913	.952	.947	.960	.937	.999	.990
.01	5*	4.877	2.076	1.145	1.100	.932	.883	.913	.903
	10	1.872	1.303	1.102	1.113	1.100	1.147	1.169	1.200
	15	1.705	1.250	1.229	1.154	1.256	1.316	1.269	1.358
	20	1.535	1.245	1.255	1.318	1.326	1.353	1.330	1.398
	25	1.543	1.312	1.286	1.358	1.253	1.427	1.450	1.441
	30	1.631	1.337	1.361	1.368	1.401	1.413	1.500	1.475

*NOTE: For shape c = 0.5, critical values correspond to sample size n = 6 instead of n = 5.

Table VIII

CRITICAL VALUES FOR THE MODIFIED CRAMER-VON MISES TEST

		Pareto Shape Parameter c							
α	n	0.5*	1.0	1.5	2.0	2.5	3.0	3.5	4.0
.20	5*	.212	.103	.078	.078	.077	.079	.082	.083
	10	.121	.083	.081	.082	.086	.088	.089	.092
	15	.108	.086	.086	.086	.090	.092	.096	.098
	20	.104	.085	.086	.091	.094	.094	.097	.099
	25	.101	.086	.090	.090	.094	.097	.101	.102
	30	.100	.086	.091	.092	.095	.097	.102	.101
.15	5*	.251	.112	.083	.084	.084	.085	.089	.092
	10	.135	.093	.090	.091	.096	.099	.099	.102
	15	.120	.094	.094	.097	.100	.103	.108	.111
	20	.115	.095	.096	.102	.105	.106	.108	.110
	25	.112	.097	.101	.102	.105	.108	.112	.115
	30	.112	.096	.102	.105	.106	.109	.116	.114
.10	5*	.304	.123	.093	.093	.093	.096	.100	.103
	10	.154	.105	.102	.104	.108	.114	.113	.119
	15	.139	.109	.109	.114	.116	.120	.126	.130
	20	.133	.109	.111	.118	.120	.123	.125	.129
	25	.130	.112	.115	.120	.121	.127	.130	.133
	30	.129	.110	.121	.121	.124	.127	.135	.131
.05	5*	.381	.139	.113	.111	.111	.112	.119	.120
	10	.184	.127	.125	.126	.131	.139	.142	.147
	15	.172	.131	.134	.140	.142	.144	.156	.161
	20	.163	.133	.136	.145	.148	.155	.153	.161
	25	.157	.139	.142	.149	.148	.157	.162	.168
	30	.159	.137	.151	.150	.156	.154	.169	.166
.01	5*	.508	.174	.157	.148	.149	.150	.157	.163
	10	.251	.191	.174	.182	.194	.199	.202	.209
	15	.255	.192	.193	.198	.207	.222	.222	.238
	20	.233	.195	.201	.212	.220	.225	.226	.239
	25	.245	.199	.206	.224	.215	.238	.240	.250
	30	.247	.202	.217	.218	.230	.243	.251	.251

*NOTE: For shape c = 0.5, critical values correspond to sample size n = 6 instead of n = 5.

Power Comparison Tables

Tables IX and X display the results of the power analysis. For sample sizes n = 5, 15, and 25, the tables indicate relative power of the K-S, A-D, and C-VM tests to reject a null hypothesis when the hypothesis claims that a random sample of data follows a Pareto distribution. For sample sizes n = 15 and 25, the power of the Chi-square test is also included. Table IX shows power values when the null hypothesized Paréto CDF has shape parameter c = 1.0. In Table X, the hypothesized shape parameter is c = 3.5. Both tables examine power performance against eight different distributions, including three variations of the Pareto distribution having different sets of parameters.

The power tables are divided into two levels of significance, α = .05 and .01. In Table IX, the first column corresponds to a Pareto distribution with shape c = 1.0. Thus, the values in the first column of Table IX approximate the level of significance α, since they represent rejection rates of the null hypothesis when H_0 is true. Similarly in Table X, the second column represents a true null hypothesis since the underlying data was generated from a Pareto distribution with shape parameter c = 3.5. Aside from these two exceptions, all other colums represent power values since they indicate rejection rates of the null hypothesis when H_0 is in fact false. A note following the tables indicates parameters of the alternate distributions.

Table IX

POWER TEST FOR THE PARETO DISTRIBUTION
H_0: Pareto Distribution at Shape c = 1.0
H_1: The data follow another distribution

Level of Significance = .05

n	Test	Par.1	Par.2	Par.3	Weibl	Gamma	Beta	Expon	Norml
		Alternate Distributions*							
5	K-S	0.046	0.061	0.050	0.288	0.123	0.227	0.074	0.311
	A-D	0.048	0.014	0.022	0.007	0.006	0.008	0.009	0.007
	CVM	0.050	0.063	0.051	0.283	0.127	0.224	0.076	0.307
15	K-S	0.048	0.145	0.107	0.979	0.657	0.933	0.290	0.979
	A-D	0.052	0.126	0.083	0.966	0.644	0.898	0.266	0.965
	CVM	0.052	0.173	0.121	0.974	0.697	0.915	0.329	0.973
	χ^2	0.043	0.118	0.086	0.860	0.480	0.738	0.235	0.878
25	K-S	0.052	0.248	0.138	1.000	0.927	1.000	0.503	1.000
	A-D	0.049	0.250	0.128	1.000	0.937	0.998	0.528	1.000
	CVM	0.050	0.256	0.143	0.999	0.926	0.996	0.504	1.000
	χ^2	0.045	0.178	0.105	0.999	0.823	0.996	0.377	0.999

Level of Significance = .01

n	Test	Par.1	Par.2	Par.3	Weibl	Gamma	Beta	Expon	Norml
5	K-S	0.010	0.021	0.021	0.171	0.067	0.115	0.034	0.172
	A-D	0.009	0.002	0.004	0.000	0.000	0.000	0.000	0.001
	CVM	0.010	0.019	0.019	0.155	0.059	0.098	0.030	0.160
15	K-S	0.015	0.059	0.035	0.941	0.448	0.852	0.150	0.937
	A-D	0.011	0.038	0.021	0.875	0.356	0.716	0.103	0.878
	CVM	0.016	0.062	0.034	0.906	0.439	0.777	0.139	0.910
	χ^2	0.006	0.031	0.016	0.645	0.172	0.400	0.064	0.669
25	K-S	0.010	0.086	0.039	0.999	0.774	0.992	0.250	0.998
	A-D	0.009	0.080	0.032	0.997	0.778	0.982	0.247	0.997
	CVM	0.010	0.100	0.046	0.997	0.792	0.982	0.274	0.998
	χ^2	0.011	0.061	0.033	0.964	0.594	0.884	0.172	0.971

* Key to Alternate Distributions:

Par.1 - Pareto (a=1, b=1, c=1)
Par.2 - Pareto (a=2, b=3, c=3.5)
Par.3 - Pareto (a=10, b=5, c=2)
Weibl - Weibull (shape = 3.5)

Gamma - Gamma (shape = 2)
Beta - Beta (P=2, Q=3)
Expon - Exponential (mean = 2)
Norml - Normal distribution

Table X

POWER TEST FOR THE PARETO DISTRIBUTION
H_0: Pareto Distribution at Shape c = 3.5
H_1: The data follow another distribution

Level of Significance = .05

n	Test	Par.1	Par.2	Par.3	Weibl	Gamma	Beta	Expon	Norml
		Alternate Distributions*							
5	K-S	0.120	0.048	0.051	0.160	0.065	0.108	0.051	0.156
	A-D	0.182	0.054	0.072	0.153	0.052	0.098	0.045	0.153
	CVM	0.122	0.051	0.050	0.212	0.074	0.148	0.055	0.208
15	K-S	0.312	0.048	0.072	0.673	0.211	0.428	0.060	0.690
	A-D	0.389	0.046	0.100	0.813	0.262	0.605	0.065	0.823
	CVM	0.332	0.043	0.080	0.814	0.278	0.602	0.076	0.826
	χ^2	0.136	0.037	0.044	0.707	0.169	0.480	0.060	0.717
25	K-S	0.472	0.045	0.086	0.928	0.387	0.763	0.084	0.942
	A-D	0.559	0.051	0.122	0.983	0.531	0.924	0.092	0.985
	CVM	0.511	0.049	0.099	0.980	0.527	0.907	0.098	0.982
	χ^2	0.245	0.036	0.048	0.940	0.317	0.784	0.071	0.948

Level of Significance = .01

n	Test	Par.1	Par.2	Par.3	Weibl	Gamma	Beta	Expon	Norml
5	K-S	0.075	0.009	0.017	0.033	0.011	0.026	0.005	0.034
	A-D	0.096	0.012	0.026	0.014	0.004	0.012	0.003	0.011
	CVM	0.064	0.011	0.014	0.053	0.015	0.041	0.010	0.056
15	K-S	0.198	0.011	0.027	0.379	0.067	0.177	0.014	0.412
	A-D	0.261	0.009	0.038	0.578	0.081	0.317	0.012	0.603
	CVM	0.224	0.011	0.031	0.609	0.101	0.347	0.017	0.630
	χ^2	0.078	0.015	0.013	0.576	0.094	0.346	0.026	0.582
25	K-S	0.341	0.010	0.040	0.794	0.167	0.479	0.021	0.807
	A-D	0.402	0.008	0.052	0.912	0.209	0.685	0.016	0.915
	CVM	0.377	0.011	0.046	0.922	0.246	0.706	0.023	0.924
	χ^2	0.130	0.009	0.011	0.878	0.148	0.614	0.022	0.882

* Key to Alternate Distributions:

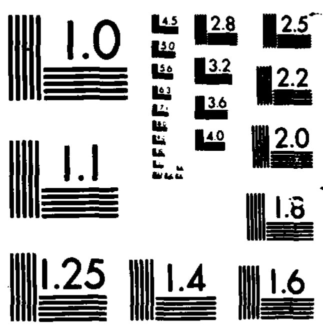
Par.1 - Pareto (a=1, b=1, c=1)
Par.2 - Pareto (a=2, b=3, c=3.5)
Par.3 - Pareto (a=10, b=5, c=2)
Weibl - Weibull (shape = 3.5)

Gamma - Gamma (shape = 2)
Beta - Beta (P=2, Q=3)
Expon - Exponential (mean = 2)
Norml - Normal distribution

AD-A163 837 MODIFIED KOLMOGOROV-SMIRNOV ANDERSON-DARLING AND CRAMER-VON MISES TESTS F..(U) AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL OF ENGI.. J E PORTER 2/2

UNCLASSIFIED DEC 85 AFIT/GSO/MA/85D-6 F/G 12/1 NL

END
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Linear Regression Tables

Tables XI and XII indicate the linear relationships existing between critical values and Pareto shape parameters. Table XI pertains to Kolmogorov-Smirnov critical values, while Table XII pertains to Cramer-von Mises critical values. No consistent linear relationship was identified for Anderson-Darling critical values.

The two tables contain linear coefficients and correlation values for each combination of sample sizes n = 10, 15, 20, 25, and 30 and levels of significance $\alpha = .20$, .15, .10, .05, and .01. No consistent linear relationship could be found for sample size n = 5. Further, the linear relationships apply only for values of the shape parameter c in the range $1.5 \leq c \leq 4.0$. Critical values for $c < 1.5$ failed to display any consistent linear trend.

Each combination of sample size and significance level has its own linear coefficients and correlation value. In each case, the relationship between critical value Y and shape parameter c is given by the simple linear regression equation $Y = b_0 + b_1c$ where b_0 corresponds to the Y-axis intercept and b_1 represents the slope of the described line. The correlation value R^2 indicates the percent of total variation explained by the regression line. Thus, R^2 is a measure of the strength of the linear relationship, with values near 1 indicating a strong linear tendency (13:250).

Table XI

COEFFICIENTS AND R^2 VALUES OF THE RELATIONSHIP*
BETWEEN KOLMOGOROV-SMIRNOV CRITICAL VALUES AND
PARETO SHAPE PARAMETERS FOR $1.5 \leq C \leq 4.0$

		Level of Significance				
n	Coeff	.20	.15	.10	.05	.01
10	b_0	.2080	.2154	.2222	.2359	.2704
	b_1	.0057	.0067	.0090	.0117	.0144
	R^2	0.998	0.997	0.993	0.997	0.993
15	b_0	.1752	.1804	.1896	.2042	.2339
	b_1	.0051	.0065	.0074	.0091	.0117
	R^2	0.977	0.993	0.999	0.990	0.987
20	b_0	.1544	.1630	.1699	.1828	.2102
	b_1	.0044	.0042	.0054	.0068	.0091
	R^2	0.973	0.969	0.964	0.960	0.935
25	b_0	.1403	.1461	.1535	.1623	.1885
	b_1	.0038	.0043	.0050	.0075	.0091
	R^2	0.980	0.991	0.963	0.994	0.964
30	b_0	.1302	.1362	.1418	.1542	.1728
	b_1	.0030	.0034	.0047	.0053	.0090
	R^2	0.944	0.947	0.946	0.967	0.979

* Relationship between K-S critical values Y and Pareto shape parameter c is approximately

$$Y = b_0 + b_1 c \quad \text{where } 1.5 \leq c \leq 4.0$$

Table XII

COEFFICIENTS AND R^2 VALUES OF THE RELATIONSHIP*
BETWEEN CRAMER-VON MISES CRITICAL VALUES AND
PARETO SHAPE PARAMETERS FOR $1.5 \leq C \leq 4.0$

		Level of Significance				
n	Coeff	.20	.15	.10	.05	.01
10	b_0	.0741	.0825	.0915	.1089	.1556
	b_1	.0045	.0050	.0067	.0095	.0137
	R^2	0.986	0.970	0.973	0.985	0.981
15	b_0	.0769	.0832	.0964	.1170	.1643
	b_1	.0053	.0069	.0083	.0106	.0178
	R^2	0.982	0.996	0.993	0.965	0.980
20	b_0	.0805	.0905	.1031	.1252	.1833
	b_1	.0047	.0051	.0065	.0089	.0135
	R^2	0.966	0.957	0.978	0.957	0.974
25	b_0	.0806	.0910	.1045	.1264	.1831
	b_1	.0055	.0059	.0072	.0102	.0166
	R^2	0.979	0.989	0.992	0.978	0.932
30	b_0	.0834	.0936	.1116	.1372	.1907
	b_1	.0047	.0055	.0054	.0074	.0161
	R^2	0.964	0.945	0.899	0.872	0.967

* Relationship between C-VM critical values Y and Pareto shape parameter c is approximately

$$Y = b_0 + b_1 c \quad \text{where } 1.5 \leq c \leq 4.0$$

Use of Tables

This section explains how to use the research results contained in Tables VI - XII.

Using Critical Value Tables. The critical values contained in Tables VI - VIII can be used to test whether a random data sample of size n = 5, 10, 15, 20, 25, or 30 follows a three-parameter Pareto distribution having specified shape parameter c = .5, 1, 1.5, 2, 2.5, 3, 3.5, or 4. Given a random sample of observed data, the following steps outline basic elements of the procedure used in testing goodness-of-fit (13:357-367):

Step 1. Determine n, the number of observations contained in the random data sample.

Step 2. Identify the null and alternative hypotheses to be tested. In this case, the hypothesized shape parameter c must also be specified. Thus, the hypotheses are:

H_0: The sample observations follow a Pareto distribution of specified shape c.

H_1: At least one of the observations does not follow the Pareto of shape c.

Step 3. Determine the desired probability of commiting a Type I error, i.e., the probability of erroneously rejecting the null hypothesis when H_0 is true. This probability is the level of significance, α (13:78).

Step 4. Order the n observations from smallest to largest.

Step 5. Assume H_0 is true and estimate the unknown location and scale parameters using an invariant estimator. If the BLUE is selected as the estimator, and the sample size is small, the estimates can be computed manually from equations (34) and (35) for c = .5, 1, or 2; equations (37) to (39) for c = 1.5; or equations (17), (18), (21), (22), and (29) for c = 2.5, 3, 3.5, or 4. For larger sample sizes, or if several samples are involved, use the FORTRAN subroutines BXVALS, BLCLE2, and BLCGT2 in Appendix A.

Step 6. Use the estimates of location a and scale b, the hypothesized shape c, and the n ordered sample observations to compute the hypothesized Pareto CDF from equation (40). Subroutine HYPCDF in Appendix A can be used if manual calculations are not practical.

Step 7. Select the type of test to be performed and compute the corresponding test statistic. Use equation (42) for the modified Kolmogorov-Smirnov test, equation (43) for the modified Anderson-Darling test, or equation (44) for the modified Cramer-von Mises test. Subroutine TESTAT in Appendix A can be used to compute test statistics for all three tests.

Step 8. Identify the critical value from Table VI, VII, or VIII, based on test type, level of significance, sample size, and hypothesized shape parameter.

Step 9. Reject the null hypothesis if the value of the test statistic exceeds the critical value. If the test statistic does not exceed the critical value, conclude that there is insufficient evidence to reject the null hypothesis (13:76).

Using Power Comparison Tables. Tables IX and X can be used to draw conclusions regarding the relative ability of a test to correctly reject a false null hypothesis. This information can then be used to select the best test for a given situation. The higher the power value, the better are the chances against commiting a Type II error because the probability of erroneously accepting a false null hypothesis is lessened (13:78).

Using Linear Regression Tables. Tables XI and XII can be used to estimate critical values for shape parameters which are not specifically listed in Tables VI and VIII, provided the hypothesized shape parameter c satisfies $1.5 \leq c \leq 4.0$. Given the sample size and specified level of significance, the linear slope and intercept values contained in Table XI can be substituted into the regression equation $y = b_0 + b_1 c$ to find the Kolmogorov-Smirnov critical value y. If the Cramer-von Mises test is involved, the values should be taken from Table XII.

Example

Suppose a maintenance unit wants to model the failure rate of a certain equipment component. Based on 10 independent random samples, the unit observes the following failure times of the component (expressed in months following initial use): 1.178, 1.127, 1.373, 1.068, 1.059, 1.010, 1.474, 4.830, 3.997, 1.799. The unit desires to test the hypothesis that the component failure times follow the Pareto distribution with shape c = 2.5. One specified requirement is that the test be designed so that the probability of erroneously rejecting a true null hypothesis must not exceed .05.

Since there are 10 random observations in the data sample, n = 10 for this example. The required level of significance is α = .05. The hypotheses are:

> H_0: The observed failure times follow the Pareto distribution of shape c = 2.5.
>
> H_1: At least one of the observations does not follow the Pareto of shape 2.5.

The next step is to arrange the random sample in ascending order: 1.010, 1.059, 1.068, 1.127, 1.178, 1.373, 1.474, 1.799, 3.997, 4.830. These values are input into subroutine BXVALS which yields B_i values of .920, .838, .754, .668, .579, .486, .389, .285, .171, and .034. These values are then input into subroutine BLCGT2, which computes the parameter estimates $\hat{a}$ = .963 and $\hat{b}$ = 1.128. Subroutine

HYPCDF is then used to compute 10 values of the hypothesized Pareto CDF: .097, .185, .201, .288, .354, .539, .608, .750, .962, and .976.

The values of n, c, and the hypothesized Pareto CDF are input into subroutine TESTAT, which computes the test statistics K-S = .162, A-D = .416, and C-VM = .042. From Table VI, the K-S critical value for α = .05, n = 10, and c = 2.5 is .265. Since the test statistic does not exceed the critical value, there is insufficient evidence to reject the null hypothesis. The same conclusion is reached from the A-D and C-VM critical values (Tables VII and VIII).

Now suppose the unit wants to test the null hypothesis that a set of n = 25 observed service times follows the Pareto distribution of shape c = 3.35. The analyst computes the K-S or C-VM test statistic as before, but the critical values are not listed for c = 3.35. Therefore, the next step is to determine the appropriate regression coefficients from Table XI or XII. For n = 25 and α = .05 the K-S coefficients are b_0 = .1623 and b_1 = .0075. The K-S critical value is $Y = b_0 + b_1 c = .1623 + .0075\ (3.35) = .1874$.

Chapter Summary

This chapter presented the results of the research conducted in response to the three objectives listed in Chapter I. Tables of critical values for the modified K-S, A-D, and C-VM tests were presented. Also included were

tables comparing powers of the K-S, A-D, and C-VM statistics against the Chi-square. Tables of regression coefficients were presented as well. The use of the tables was explained, and an example was described.

The research results are further analysed and discussed in the next chapter.

VI. ANALYSIS AND DISCUSSION

Chapter Overview

This chapter discusses the results presented in Chapter V. Observations are made concerning the tables of critical values, power comparisons, and regression coefficients, including an explanation as to how the computer programs were verified and validated.

Critical Values

The critical value tables generated for this thesis are located in Chapter V. For the K-S test (Table VI), the critical values for a given level of significance and shape parameter decrease as the sample size increases. Further, the size of the decrease becomes smaller at larger values of n. This trend suggests that the K-S critical values may converge to a lower limit as the sample size increases. However, the use of sample sizes larger than 30 would have required much more computer processing time, and thus was beyond the scope of this thesis. The A-D critical values (Table VII) exhibit a different pattern. The values for each combination of significance level and shape parameter generally decrease from n = 5 to 20 and increase from n = 20 to 30, suggesting a convergence between 15 and 20. Similarly, the C-VM critical values (Table VIII) appear to converge between n = 25 and 30,

since the values consistently decrease until n = 30, then begin to increase.

An important observation is made when the table of modified K-S values is compared to a standard (unmodified) K-S table (13:462). For each value of n in Table VI, the critical values for shape 1 or 2 at a .05 significance level are nearly the same as the critical values for a .20 significance level using the standard table. Thus the result of using the standard K-S table when location and scale parameters are estimated would be to obtain an extremely conservative test in the sense that the actual significance level would be much lower than that given by the standard table.

Power Comparison

The power comparison tables generated for this thesis are located in Chapter V. Values in Table IX pertain to a null hypothesis for which the Pareto shape parameter is 1.0, whereas in Table X the hypothesized shape parameter is 3.5. Both tables are divided into two sections based on a level of significance of .05 or .01. It is obvious from the tables that none of the three tests developed in this thesis is very powerful when the sample size is only five. Nevertheless, they at least provide some means of testing goodness-of-fit for sample sizes which are too small to use the Chi-square test. For sample sizes of 15 or 25, the powers improve dramatically.

For each alternative distribution the three tests tended to be more powerful than the Chi-square. Two sets of Chi-square critical values were examined. The first set of values was taken from a standard table of Chi-square critical values corresponding to 2 degrees of freedom (13:432). After completing 5000 Monte Carlo repetitions, it was discovered that the tabled Chi-square value for a level of significance of .05 displayed a probability of a Type I error (i.e., rejecting H_0 when true) of .10, which was twice the claimed level of significance of .05. Similarly, the probability of Type I error for a claimed level of significance .01 was, in fact, .02. This discrepancy was due to the fact that the tabled Chi-square values represent only an approximation of the actual asymptotic distribution of the Chi-square, so that the actual value lies somewhere between Chi-square with 2 degrees of freedom and Chi-square with 4 degrees of freedom (34:401-402). Since the Type I errors were twice their expected value, a second set of Chi-square critical values was generated using Monte Carlo simulation in the same manner as was used to generate critical values for the K-S, A-D, and C-VM tests. As apparent from Tables IX and X, the second set of Chi-square values display Type I error rates which is much closer to the claimed level of significance of .05 or .01. Therefore, these values were used in the power comparison tables rather than the less accurate values stemming from the standard Chi-square table.

The modified K-S, A-D, and C-VM tests are especially powerful when the sample data are taken from the Weibull, the Beta, or the normal distribution. On the other hand, the three tests display relatively low power in their ability to distinguish against the exponential or the Pareto with different shape parameters. In general, the K-S test has higher power tham the others when the hypothesized shape parameter is 1.0. When the shape parameter is 3.5, the C-VM test tends to be more powerful. Next to the Chi-square, the A-D test appears to have the lowest power in most cases.

Regression Analysis

The regression tables generated for this thesis are also located in Chapter V. Table XI contains regression coefficients and correlation values for the modified Kolmogorov-Smirnov test, while Table XII contains regression values for the Cramer-von Mises test.

It is apparent from Tables VI and VIII that for a given significance level and sample size except n = 5, the K-S and C-VM critical values decrease from shape parameter 0.5 to 1.5, then steadily increase for shapes 1.5 to 4.0. Using the method of least squares, a simple linear regression analysis was performed on the critical values. The correlation of regression on the shape parameter interval 0.5 to 1.5 was in most cases less than .80. However, the regression relationships on the shape interval 1.5 to 4.0 showed very

strong correlation (.97 or higher in most cases). Therefore, regression coefficients corresponding to the interval $1.5 \leq c \leq 4.0$ were included in Tables XI and XII.

No consistent linear trend could be identified for the Anderson-Darling critical values. In general the values seem to decrease on the interval $0.5 \leq c \leq 2.5$ and then increase on $2.5 \leq c \leq 4.0$. However, when least squares regression was applied to the two intervals, the correlation values tended to be less than .80 in most cases. Therefore, it was decided not to include a regression table for the A-D test.

Verification and Validation

The critical values were computed by the CRITVAL program and associated subroutines contained in Appendix A. The power study was conducted using the POWER program and subroutines in Appendix B. The purpose of verification was to ensure that the concepts and equations developed in this thesis were reflected accurately in the computer code. The five verification techniques suggested by Banks and Carson (5:379) were implemented as follows:

1. Have the code checked. The code was checked by two individuals knowlegeable of FORTRAN programming. One of the individuals, Charek, was also very familiar with the logic required for computing parameter estimates for the Pareto distribution, since he too has conducted extensive research in this area (12).

2. Make a flow diagram. Flow diagrams illustrating the logic involved in generating critical values served as the basis of the program and were closely followed during the actual writing of the program. The diagrams are included in Appendices A and B.

3. Examine a wide variety of output. The output of each subroutine and the results of each individual computation was checked through extensive use of print statements. Each computational stage was checked at least once against manual calculations to ensure the expected values were produced. A pre-production run involving 50 replications was thoroughly examined for reasonableness prior to the final production run of 5000 replications.

4. Print the input parameters. During the test runs, input parameters were printed before and after each calculation to ensure against any inadvertant alteration of parameters.

5. Make the code self-documenting. Extensive comments have been incorporated into the programs and subroutines to allow easy interpretation of the logic. At the beginning of each program component, every variable is defined and the purpose explained.

Validation of the computer programs was provided in the results of the power study. For each hypothesized shape parameter and sample size, the K-S, A-D, and C-VM tests

displayed a Type I error rate equal to or very near the claimed level of significance. This fact validates the critical values as well as the power comparison values.

Chapter Summary

The results of this thesis are presented in Tables VI-XII. The results of the power study show that the three tests developed in this thesis offer tests which can be used with small sample sizes and are more powerful than the Chi-square at larger sample sizes. The programs used to generate the tables were thoroughly verified and validated.

Conclusions and recommendations for further study are presented in the next chapter.

VII. CONCLUSIONS AND RECOMMENDATIONS

Conclusions

The following conclusions are based on the results contained in this thesis:

1. The first research objective listed in Chapter I has been successfully fulfilled. Tables VI-VIII contain critical values of the modified Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Cramer-von Mises (C-VM) tests. The validity of these critical values has been verified by a Monte Carlo power study which has shown that all three tests achieve the claimed level of significance when the null hypothesis is true. Therefore, each table of critical values can be used to test whether a random sample of data follows the three-parameter Pareto distribution with specified shape parameter.

2. The second research objective has also been completed successfully. The results of the power study are contained in Tables IX and X. It appears that none of the three tests developed in this thesis is very powerful when the sample size is only five. For sample sizes of 15 or 25, however, the powers improve dramatically. For each of the alternative distributions considered, the three tests tended to be more powerful than the Chi-square, as expected. The three tests are especially powerful when the sample data are

taken from the Weibull, the Beta, or the normal distribution. In general, the K-S test has higher power than the others when the hypothesized shape parameter is 1.0. When the shape parameter is 3.5, the C-VM test tends to be more powerful. Next to the Chi-square, the A-D test appears to have the lowest power in most cases.

3. Successful completion of the third research objective has revealed a strong linear relationship between shape parameters and critical values for the K-S and C-VM tests. Linear coefficients and correlation values are contained in Tables XI and XII. However, no consistent functional relationship could be identified for the A-D test.

Recommendations

Based on observations made during the investigation for this thesis, the following research areas are proposed for further study:

1. Apply the techniques used in this thesis to generate modified K-S, A-D, and C-VM tests for other distribution functions.

2. Investigate whether other types of goodness-of-fit tests can be modified through Monte Carlo techniques. For example, if the S statistic of Mann, Scheuer, and Fertig (38) can be modified for the Pareto distribution, a power study can be conducted to determine whether the S statistic is more powerful than the K-S, A-D, or C-VM tests.

3. Derive the maximum likelihood estimators of location and scale for the three-parameter Pareto.

4. Compute critical values for sample sizes and Pareto shape parameters not specifically included in Tables VI-VIII. For example, the tables can be expanded to include all sample sizes from 3 to 100 and shape parameters from 0.25 to 10.

5. Increase the accuracy of the critical values by using various techniques (5:406-442) of experimental design (e.g., increased repetitions, multiple batch runs, replications, antithetic random number seeds, analysis of variance, etc.) to reduce the inherent uncertainty and to determine the amount of variance involved.

6. Apply more sophisticated regression techniques to determine the functional relationship between Pareto shape parameters and Anderson-Darling critical values.

7. Apply the results of this thesis to earlier studies (Chapter III) involving the Pareto distribution. For example, Berger and Mandelbrot's (7) conclusion that the Pareto can be used to model errors in communications circuits can now be tested for goodness-of-fit.

8. Further investigate potential applications of the Pareto distribution as an accurate model of actual phenomena. The tests developed in this thesis contribute to the usefulness of the Pareto distribution which, in many situations, should be considered as a viable model when simulating or testing the underlying distribution of a given population.

APPENDIX A

Computer Program and Subroutines

for Generating Critical Values

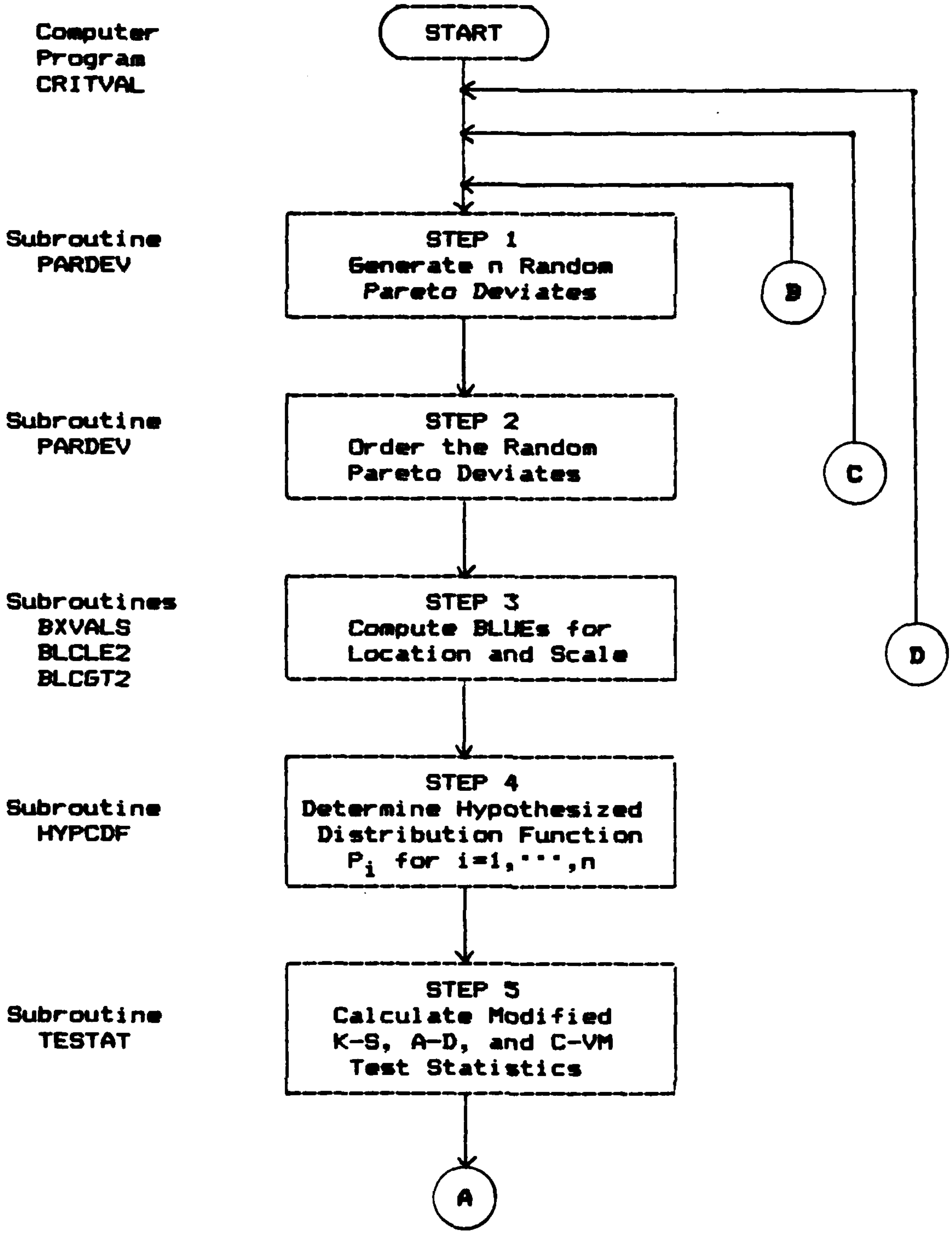


Fig 6. Procedure for Generating Critical Values

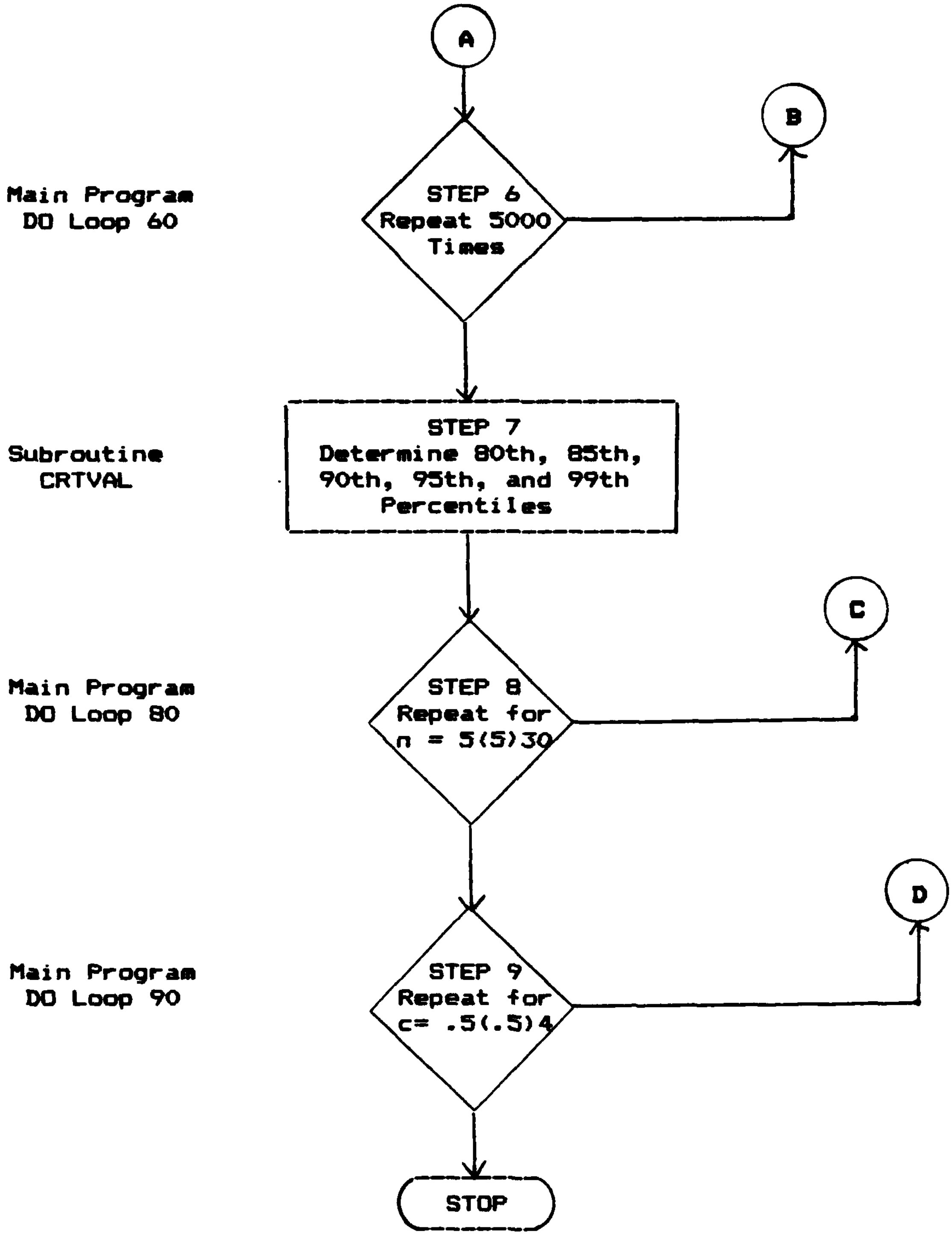


Fig 6 (Continued). Procedure for Generating Critical Values

```
c***** Classroom Support Computer (CSC) - VAX 11/785 - VMS 4.1 ****
c
c******  CRITVAL  PROGRAM FOR PARETO GOODNESS-OF-FIT TESTS   ******
c
c*****************************************************************
c*****************************************************************
c**                                                             **
c**   B E G I N   C R I T V A L   M A I N   P R O G R A M   **
c**                                                             **
c*****************************************************************
c
c  Ref: Appendix A, Figure 6.
c
c=================================================================
c
c  Purpose:
c            1.  Generate critical value tables for the modified
c            Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D),
c            and Cramer-von Mises (C-VM) tests for the three-
c            parameter Pareto distribution when location and
c            scale parameters must be estimated from sample data.
c            2.  Provide extensive commentary to help novice prog-
c            rammers develop similar goodness-of-fit programs.
c            Thus, diagnostic print routines have been retained as
c            part of the commentary rather than deleted.
c
c=================================================================
c
c  Variables:
c        dseed = random number seed
c            c = shape parameter
c            n = sample size
c         nshp = shape parameter counter (8 different values)
c         nsiz = sample size counter (6 different values of n)
c         npct = percentile counter (5 different percentiles)
c          nst = number of test statistics to be used
c           it = iteration counter (5000 repetitions required)
c           KS = array of values of modified K-S test statistic
c          CVM = array of values of modified C-VM test statistic
c           AD = array of values of modified A-D test statistic
c        alpha = level of significance
c
c=================================================================
c
c  Input:
c       nst = number of repetitions (input at computer terminal)
c     dseed = random number seed (input at computer terminal)
c
c=================================================================
```

```
c
c  Subroutines:
c
c    PARDEV - Generates n ordered Pareto deviates
c    BXVALS - Calculates B values and summations of B and Bx
c    BLCLE2 - Finds BLUEs for location and scale when c <= 2
c    BLCGT2 - Finds BLUEs for location and scale when c > 2
c    HYPCDF - Computes the Hypothesized Pareto CDF
c    TESTAT - Calculates the K-S, A-D, and C-VM test statistics
c    CRTVAL - Determines critical values from plotting positions
c
c==================================================================
c
c  Calculate:
c
c      nc = n * c
c
c      Plotting Positions (Eqn 51):
c
c         Y(i) = (i - 0.3)/(nst + 0.4) for i = 1,....,nst(=5000)
c
c==================================================================
c
c  Output:
c
c    KScrit = 3-D array of critical values for modified K-S test
c    ADcrit = 3-D array of critical values for modified A-D test
c    CVcrit = 3-D array of critical values for modified C-VM test
c
c=================================================================
c
c  Declare Variables:
c
       common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
     1        Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
     1        KScrit,ADcrit,CVcrit,Y
       integer n,nsiz,nshp,it,npct,nst
       real x(30),ablu,bblu,B(30),D,KS(5000,6,8),AD(5000,6,8),
     1      CVM(5000,6,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,P(30),
     1      KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),r(30),pct,
     1      Y(5002),alpha
       double precision dseed
c
c     ** Open Output Files to Store Computed Critical Values: **
            open (unit=7,file='CRIT',status='new')
c
c     **  Number of Test Statistics to be Used on Each Run:  **
            print*,'The Monte Carlo analysis will require'
            print*,'     5000 test statistics.'
            print*,'Enter the number to be used for this run:'
            read*,nst
c
```

```
c      **   Calculate 5002 Plotting Positions on the Y-axis:   **
c
              Y(0) = 0.0
          do 10 i = 1,nst
              Y(i) = (i - 0.3)/(nst + 0.4)
  10      continue
              Y(nst + 1) = 1.0
c
           print*,' '
           print*,'SELECTED MEDIAN RANKS PLOTTING POSITIONS'
           print*,'  TO BE USED TO FIND CRITICAL VALUES:'
           print*,' '
           print*,'        Y(5001) = ',Y(5001)
           print*,'        Y(5000) = ',Y(5000)
           print*,'99PCT: Y(4950) = ',Y(4950)
           print*,'95PCT: Y(4750) = ',Y(4750)
           print*,'90PCT: Y(4500) = ',Y(4500)
           print*,'85PCT: Y(4250) = ',Y(4250)
           print*,'80PCT: Y(4000) = ',Y(4000)
           print*,'        Y(0001) = ',Y(1)
           print*,'        Y(0000) = ',Y(0)
           print*,'========================================'
c
c      **     Plotting Positions Computation Complete       **
c
      print*,'Enter random number seed or "1." for default:'
      read*,dseed
            if (dseed .eq. 1.) dseed = 123457.d00
      print*,' '
      print*,'STANDBY . . . COMPUTATIONS IN PROGRESS'
c
      nshp = 0
c
c --- Begin DO Loop 90 for Shape Parameter Values c = .5(.5)4 ---
c
      do 90 shape = 0.5,4.0,.5
            c = shape
          nshp = nshp + 1
c
c        Write Headings for Output Data:
            write(7,52)
            write(7,51)
            write(7,52)
            write(7,54)
            write(7,52)
            write(7,56)
c
            nsiz = 0
c
c      ---  Begin DO Loop 80 for Sample Sizes n = 5(5)30  ---
c
             do 80 nsamp = 5,30,5
```

```
               if ( (c.eq.0.5) .and. (nsamp.eq.5) ) then
c              the BLUEs do not exist, so we must let:
                  n = 6
               else
                  n = nsamp
               end if
c
               nsiz = nsiz + 1
                 nc = n * c
c
               write(7,58)
c
c           --- Begin DO Loop 60 for 5000 Iterations ---
c
               do 60 it = 1,nst
c
c              ** Perform Steps 1 & 2 of Fig 6: **
c
                      call PARDEV
c
c              **  Perform Step 3 of Figure 6:   **
c
                      call BXVALS
c
                      if (c .le. 2.0) then
                         call BLCLE2
                      else
                         call BLCGT2
                      end if
c
c              **  Perform Step 4 of Figure 6: **
c
                      call HYPCDF
c
c              **  Perform Step 5 of Figure 6: **
c
                      call TESTAT
c
   60          continue
c
c           ---  End DO Loop 60 for 5000 Iterations  ---
c              ** Completes Step 6 of Figure 6 **
c
c              **  Perform Step 7 of Figure 6: **
c
c           ---   Begin DO Loop 70 for Percentiles   ---
c
         do 70 npct = 1,5
c
c
                      call CRTVAL
```

```
c
c                  --  Write CRTVAL Output to File --
                write(7,62),1.-pct,n,c,KScrit(nsiz,nshp,npct),
     1             ADcrit(nsiz,nshp,npct),CVcrit(nsiz,nshp,npct)
c
                  print*,' '
                  print*,' CRITICAL VALUES FROM MAIN PROGRAM'
                  print*,' pct =',pct,' n=',n,' ** c=',c
                  print*,' K-S =',KScrit(nsiz,nshp,npct),
     1                  ' A-D =',ADcrit(nsiz,nshp,npct),
     1                  ' CVM =',CVcrit(nsiz,nshp,npct)
                  print*,' '

  70      continue
c
c            ---     End DO Loop 70 for Percentiles   ---
c
  80          continue
c
c       ---  End DO Loop 80 for Sample Sizes n = 5(5)30   ---
c                  ** Completes Step 8 of Figure 6 **
c
  90  continue
c
c ---  End DO Loop 90 for Shape Parameter Values c = .5(.5)4 ---
c                  ** Completes Step 9 of Figure 6 **
c
c*****************************************************************
c
c   OUTPUT INSTRUCTIONS:  The remainder of the main program
c   consists of commands to format the output data and write
c   the data and headers to a file which can be printed out
c   in hardcopy.
c
c*****************************************************************
c
c *** Write KS Critical Value Tables to File by Alpha Level: ***
c
      write(7,52)
      write(7,130)
      write(7,52)
      write(7,132)
      write(7,52)
      write(7,200)
      write(7,201)
      write(7,52)
c
      npct = 0
c
c ---Begin DO Loop 105 to Sort Critical Values by Alpha Level---
c
      do 105 npct = 1,5
```

```
      c
             if (npct .ne. 5) alpha = .25 - (.05*npct)
             if (npct .eq. 5) alpha = .01
      c
             nsiz = 0
             n = 0
      c
      c    --- Begin DO Loop 107 to Sort Output by Sample Size ---
      c
             do 107 nsiz = 1,6
      c
                n = 5 * nsiz

                Write(7,120),alpha,n,KScrit(nsiz,1,npct),KScrit
          1        (nsiz,2,npct),KScrit(nsiz,3,npct),KScrit(nsiz,
          1        4,npct),KScrit(nsiz,5,npct),KScrit(nsiz,6,npct),
          1        KScrit(nsiz,7,npct),KScrit(nsiz,8,npct)
      c
       107   continue
      c
      c    ---   End DO Loop 107 After Sorting by Sample Size  ---
      c
             write(7,201)
      c
       105  continue
      c
      c ---   End DO Loop 105 After Sorting Output by Alpha Level  ---
      c
      c
      c *** Write AD Critical Value Tables to File by Alpha Level: ***
      c
            write(7,52)
            write(7,140)
            write(7,52)
            write(7,142)
            write(7,52)
            write(7,200)
            write(7,201)
            write(7,52)
      c
            npct = 0
      c
      c ---Begin DO Loop 115 to Sort Critical Values by Alpha Level---
      c
            do 115 npct = 1,5
      c
               if (npct .ne. 5) alpha = .25 - (.05*npct)
               if (npct .eq. 5) alpha = .01
      c
               nsiz = 0
               n = 0
      c
```

```
      c     --- Begin DO Loop 117 to Sort Output by Sample Size ---
      c
                do 117 nsiz = 1,6
      c
                   n = 5 * nsiz

                   Write(7,120),alpha,n,ADcrit(nsiz,1,npct),ADcrit
           1          (nsiz,2,npct),ADcrit(nsiz,3,npct),ADcrit(nsiz,
           1          4,npct),ADcrit(nsiz,5,npct),ADcrit(nsiz,6,npct),
           1          ADcrit(nsiz,7,npct),ADcrit(nsiz,8,npct)
      c
       117      continue
      c
      c     ---   End DO Loop 117 After Sorting by Sample Size  ---
      c
                write(7,201)
      c
       115  continue
      c
      c ---   End DO Loop 115 After Sorting Output by Alpha Level  ---
      c
      c
      c *** Write CVM Critical Value Tables to File by Alpha Level ***
      c
              write(7,52)
              write(7,150)
              write(7,52)
              write(7,152)
              write(7,52)
              write(7,200)
              write(7,201)
              write(7,52)
      c
             npct = 0
      c
      c ---Begin DO Loop 125 to Sort Critical Values by Alpha Level---
      c
             do 125 npct = 1,5
      c
                if (npct .ne. 5) alpha = .25 - (.05*npct)
                if (npct .eq. 5) alpha = .01
      c
                nsiz = 0
                n = 0
      c
      c     --- Begin DO Loop 127 to Sort Output by Sample Size ---
      c
                do 127 nsiz = 1,6
      c
                   n = 5 * nsiz

```

```
                Write(7,120),alpha,n,CVcrit(nsiz,1,npct),CVcrit
     1             (nsiz,2,npct),CVcrit(nsiz,3,npct),CVcrit(nsiz,
     1             4,npct),CVcrit(nsiz,5,npct),CVcrit(nsiz,6,npct),
     1             CVcrit(nsiz,7,npct),CVcrit(nsiz,8,npct)
c
 127     continue
c
c    ---   End DO Loop 127 After Sorting by Sample Size  ---
c
         write(7,201)
c
 125  continue
c
c ---   End DO Loop 125 After Sorting Output by Alpha Level  ---
c
c      Specify Format for Hardcopy Output Data and Headers:
c
  51  format(' *********************************************')
  52  format(' ')
  54  format(' ** PARETO CRITICAL VALUES FOR SHAPE C =      **')
  56  format(' alpha',3X,'n',4X,'c',7X,'KS',8X,'AD',8X,'CVM')
  58  format(' ---------------------------------------------')
  62  format(' ',T3,F3.2,I5,F6.1,3F10.4)
 120  format(' ',T3,F3.2,I5,F8.3,7F9.3)
 130  format('1',36X,'Table VI')
 132  format(20X,'CRITICAL VALUES FOR THE MODIFIED K-S TEST')
 140  format('1',36X,'Table VII')
 142  format(20X,'CRITICAL VALUES FOR THE MODIFIED A-D TEST')
 150  format('1',35X,'Table VIII')
 152  format(19X,'CRITICAL VALUES FOR THE MODIFIED C-VM TEST')
 200  format(' alpha',3X,'n',4X,'c=.5',5X,'1.0',6X,'1.5',6X,
     1         '2.0',6X,'2.5',6X,'3.0',6X,'3.5',6X,'4.0')
 201  format(81('-'))
c
      close(7)
c
      end
c
c===================================================================
c                          END MAIN PROGRAM
c*******************************************************************
```

```
      Subroutine PARDEV
c***********************************************************************
c**                                                                   **
c**      B E G I N    S U B R O U T I N E    P A R D E V              **
c**                                                                   **
c***********************************************************************
c
c  Ref:  Appendix A, Fig 6, Steps 1 & 2.
c
c=======================================================================
c
c  Purpose:  For a specified sample size n, generate n random
c            deviates from a Pareto distribution with location and
c            scale parameters set to one (a = b = 1) and the shape
c            parameter c set to some specified positive value.
c
c=======================================================================
c
c  Variables:
c                  r = array containing n random numbers
c                  c = shape parameter
c                  x = array containing n Pareto deviates
c                  n = sample size
c              dseed = random number seed
c
c=======================================================================
c
c  Input:      dseed = random number seed  (from MAIN program)
c                  c = shape parameter = .5(.5)4  (MAIN DO Loop 90)
c                  n = sample size = 5(5)30  (MAIN DO Loop 80)
c
c=======================================================================
c
c  IMSL Subroutines:
c
c   GGUBS - generates random numbers uniformly distributed on (0,1)
c   VSRTA - arranges a set of numbers in ascending order
c
c=======================================================================
c
c  Calculate:
c
c     x(j) = (1/r(j)) ** (1/c)   for j = 1,2,...,n  (from eqn 48)
c
c=======================================================================
c
c  Output:      x = array of n ordered Pareto deviates
c
c=======================================================================
c
c  Declare Variables:
c
```

```
      common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
     1       Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
     1       KScrit,ADcrit,CVcrit,Y
      real x(30),ablu,bblu,B(30),D,KS(5000,6,8),AD(5000,6,8),
     1     CVM(5000,6,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,P(30),
     1     r(30),KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),
     1     Y(5002)
      integer n,npct
      double precision dseed
c
c--- Begin DO Loop 10 to Generate n Random Pareto Deviates ---
c
      do 10 j = 1,n
c
c         Use IMSL subroutine to generate random numbers:
            call GGUBS(dseed,n,r)
c
c         Use eqn 48 to transform them to Pareto deviates:
            x(j) = (1.0/r(j))**(1.0/c)
c
   10 continue
c
c---   End DO Loop 10 after Generating n Random Deviates  ---
c              **  (Completes Step 1 of Figure 6)  **
c
c  Use IMSL subroutine to place the deviates in ascending order:
      call vsrta(x,n)
c              **  (Completes Step 2 of Figure 6)  **
c
      return
      end
c
c=================================================================
c                        END SUBROUTINE PARDEV
c*****************************************************************
```

```
      Subroutine BXVALS
c******************************************************************
c**                                                             **
c**      B E G I N   S U B R O U T I N E   B X V A L S          **
c**                                                             **
c******************************************************************
c
c  Ref:  Appendix A, Fig. 6, Step 3.
c
c=================================================================
c
c  Purpose:  For a given sample size n, calculate the B values
c            used to find the BLUEs of location and scale.  Also
c            find the sum of the first n-1 values of B(i).  Then,
c            compute the three values equal to the sums of the
c            first n-1, the first n-2, and (for c = .5, 1, or 2)
c            the first n -2/c values of B(i)x(i).
c
c=================================================================
c
c  Variables:   c = shape parameter
c               n = sample size
c               x = array containing n ordered Pareto deviates
c               B = array containing n values of B
c           Bsum1 = sum of B(i) values for i = 1,2,...,(n-1)
c          Bxsum1 = sum of B(i)x(i) for i = 1,2,...,(n-1)
c          Bxsum2 = sum of B(i)x(i) for i = 1,2,...,(n-2)
c          Bxsm2c = sum of B(i)x(i) for i = 1,2,...,(n-2/c)
c
c=================================================================
c
c  Input:     c = shape parameter = .5(.5)4  (from MAIN DO Loop 90)
c             n = sample size = 5(5)30  (from MAIN DO Loop 80)
c            nc = n*c   (from MAIN program)
c             x = ordered Pareto deviates (from PARDEV)
c
c=================================================================
c
c  Calculate:
c
c         B(i) = [1 - 2/c(n-i+1)] * B(i-1)      (eqn 29)
c
c        Bsum1 = B(1) + B(2) + ... + B(n-1)
c
c       Bxsum1 = B(1)*x(1) + ... + B(n-1)*x(n-1)
c
c       Bxsum2 = B(1)*x(1) + ... + B(n-2)*x(n-2)
c
c       Bxsm2c = B(1)*x(1) + ... + B(n-2/c)*x(n-2/c)
c
c=================================================================
c
```

```
c  Output:
c            B = array containing n values of B
c        Bsum1 = sum of first (n-1) B values
c       Bxsum1 = sum of first (n-1) B*x values
c       Bxsum2 = sum of first (n-2) B*x values
c       Bxsm2c = sum of first (n-2/c) B*x (if 2/c is integer)
c
c==================================================================
c
c  Declare Variables:
c
      common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
     1       Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
     1       KScrit,ADcrit,CVcrit,Y
      real x(30),ablu,bblu,B(30),D,KS(5000,6,8),AD(5000,6,8),
     1     CVM(5000,6,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,P(30),
     1     KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),Y(5002)
      integer n
      double precision dseed
c
c  Calculate the first B value (eqn 25):
c
      B(1) = 1.0 - 2.0/nc
c
c --- Begin DO Loop 10 to Find the 2nd thru nth B values ---
c
      do 10 j = 2,n
         B(j) = B(j-1) * (1.0 - (2.0/(c*(n-j+1))))
 10   continue
c
c     --- End DO Loop 10 ---
c
      Bsum1 = 0
c
c --- Begin DO Loop 20 to Sum the First n-1 Values of B  ---
c
      do 20 k=1,(n-1)
         Bsum1 = Bsum1 + B(k)
 20   continue
c
c     --- End DO Loop 20 ---
c
      Bxsum1 = 0
c
c --- Begin DO Loop 30 to Sum the First n-1 Values of Bx ---
c
      do 30 l=1,(n-1)
         Bxsum1 = Bxsum1 + (B(l)*x(l))
 30   continue
c
c     --- End DO Loop 30 ---
c
```

```
      Bxsum2 = Bxsum1 - (B(n-1)*x(n-1))
c
c --- Find Bxsm2c When 2/c is an Integer (c=.5, 1, or 2) ---
c
      Bxsm2c = 0
c
      if (c .eq. 1.0) then
         Bxsm2c = Bxsum2
      else if (c .eq. 2.0) then
         Bxsm2c = Bxsum1
      else if (c .eq. 0.5) then
         Bxsm2c = Bxsum2 - (B(n-3)*x(n-3)) - (B(n-2)*x(n-2))
      end if
c
      return
      end
c
c=================================================================
c                              END SUBROUTINE BXVALS
c*****************************************************************
```

```
      Subroutine BLCLE2
c*******************************************************************
c**                                                               **
c**        B E G I N   S U B R O U T I N E   B L C L E 2          **
c**                                                               **
c*******************************************************************
c
c  Ref:  Appendix A, Figure 6, Step 3.
c
c===================================================================
c
c  Purpose:  Given an ordered sample of size n and specified shape
c            c<=2, calculate the BLUEs of location a and scale b.
c
c===================================================================
c
c  Variables:
c              x = array containing n ordered Pareto deviates
c              c = shape parameter
c              n = sample size
c              B = array of B values used to calculate the BLUEs
c             nc = product of n and c
c          Coef1 = coefficient used to compute BLUE of location a
c          Coef2 = coefficient used to compute BLUE of location a
c          Coef3 = coefficient used to compute BLUE of scale b
c         Bxsum2 = sum of B(i)*x(i) terms for i = 1,...,n-2
c         Bxsm2c = sum of B(i)*x(i) terms for i = 1,...,n-2/c
c           ablu = BLUE of the location parameter a
c           bblu = BLUE of the scale parameter b
c              U = value used to compute BLUEs when c = 1.5
c          Termi = terms used to compute U (i=1,2,3)
c
c===================================================================
c
c  Input:      x = array of n ordered Pareto deviates (from PARDEV)
c              c = shape parameter = 0.5, 1.0, 1.5, or 2.0
c              n = sample size = 5(5)30  (from MAIN DO Loop 80)
c             nc = n*c (from MAIN program)
c              B = array containing n values of B (from BXVALS)
c         Bxsum2 = sum of first n-2 values of B (from BXVALS)
c         Bxsm2c = sum of first n-2/c values of B (from BXVALS)
c
c===================================================================
c
c  Calculate (if c = 0.5, 1, or 2):
c
c          Coef1 = [(c+1)*(c+2)] / [(nc-2)*(nc-c-2)]
c          Coef2 = (nc-2) / (c+2)
c
c           ablu = x(1) - Coef1 * [Bxsm2c - (Coef2*x(1))]  (eqn 34)
c           bblu = (nc-1) * [x(1) - ablu]     (eqn 35)
c
```

```
c--------------------------------------------------------------------
c  Calculate (if c = 1.5):
c
c          Term1 = (nc-2) * (nc-c-2)
c          Term2 = nc * (c-2) * B(n-1)
c          Term3 = (nc-1) * (c+2)
c          Coef3 = [(nc-1)/nc] * (nc-2-U)
c              U = (Term1 - Term2) / Term3    (eqn 39)
c
c           ablu = x(1) - bblu / (nc-1)     (eqn 37)
c           bblu = (1/U) * [(c+1)*(Bxsum2) + (2c-1)*B(n-1)*x(n-1)
c                  - Coef3 * x(1)]     (eqn 38)
c====================================================================
c  Output:
c           ablu = BLUE of location parameter a
c           bblu = BLUE of scale parameter b
c====================================================================
c  Declare Variables:
c
        common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
     1         Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
     1         KScrit,ADcrit,CVcrit,Y
        integer n
        real x(30),ablu,bblu,B(30),D,c,nc,Bsum1,Bxsum1,Bxsum2,
     1       Bxsm2c,P(30),Term1,Term2,Term3,Coef1,Coef2,Coef3,U,
     1       KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),Y(5002)
        double precision dseed
c
        if ((c.eq.0.5) .or. (c.eq.1.0) .or. (c.eq.2.0)) then
            Coef1 = ((c+1.0)*(c+2.0)) / ((nc-2.0)*(nc-c-2.0))
            Coef2 = (nc-2.0) / (c+2.0)
             ablu = x(1) - Coef1 * (Bxsm2c - (Coef2*x(1)))
             bblu = (nc-1.0) * (x(1) - ablu)
c
        else if (c .eq. 1.5) then
            Term1 = (nc-2.0) * (nc-c-2.0)
            Term2 = nc * (c-2.0) * B(n-1)
            Term3 = (nc-1.0) * (c+2.0)
                U = (Term1 - Term2) / Term3
            Coef3 = ((nc-1.0)/nc) * (nc-2.0-U)
             bblu = (1.0/U) *( (c+1.0) * (Bxsum2)
     1            + (2.0*c-1.0)*B(n-1)*x(n-1) - Coef3 * x(1) )
             ablu = x(1) - (bblu / (nc-1.0))
c
        end if
c
        return
        end
c
c===================================================================
c                       END SUBROUTINE BLCLE2
c*******************************************************************
```

```
      Subroutine BLCGT2
c*****************************************************************
c**                                                             **
c**       B E G I N    S U B R O U T I N E    B L C G T 2       **
c**                                                             **
c*****************************************************************
c
c  Ref:  Appendix A, Figure 6, Step 3.
c
c=================================================================
c
c  Purpose:  Given an ordered sample of size n and a specified
c            shape c > 2, calculate the best linear unbiased
c            estimates (BLUEs) of location and scale.
c
c=================================================================
c
c  Variables:  x = array containing n ordered Pareto deviates
c              c = shape parameter
c              n = sample size
c             nc = product of n and c
c              B = array of B values used to calculate the BLUEs
c          Bsum1 = sum of B(i) terms for i = 1,...,n-1
c         Bxsum1 = sum of B(i)*x(i) terms for i = 1,...,n-1
c              D = value used to calculate the BLUEs
c             YV = value used to calculate the BLUEs
c           ablu = BLUE for location parameter a
c           bblu = BLUE for scale parameter b
c
c==================================================================
c
c  Input:      x = array of ordered Pareto deviates (from PARDEV)
c              c = shape parameter = 2.5, 3.0, 3.5, or 4.0
c              n = sample size = 5(5)30  (from MAIN DO Loop 80)
c             nc = n*c  (from MAIN Program)
c              B = array of B values (from BXVALS)
c          Bsum1 = sum of first (n-1) B values (from BXVALS)
c         Bxsum1 = sum of first n-1 B*x values (from BXVALS)
c
c==================================================================
c
c  Calculate:
c
c           D = [(c+1) * Bsum1] + [(c-1) * B(n)]          (eqn 21)
c
c         YV  = (c+1)*Bxsum1 + (c-1)*B(n)*x(n) - D*x(1)   (eqn 22)
c
c        ablu = x(1) - YV/[(nc-1)*(nc-2) - D*nc]           (eqn 17)
c
c        bblu = (nc-1) * [ x(1) - ablu ]                  (eqn 18)
c
c==================================================================
```

```
c
c  Output:     ablu = BLUE for location a
c              bblu = BLUE for scale b
c
c==================================================================
c
c  Declare Variables:
c
         common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
     1        Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
     1        KScrit,ADcrit,CVcrit,Y
         integer n
         real x(30),ablu,bblu,B(30),D,KS(5000,6,8),AD(5000,6,8),YV,
     1      CVM(5000,6,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,P(30),
     1      KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),r(30),
     1      Y(5002)
         double precision dseed
c
           D = ((c+1.0) * Bsum1) + ((c-1.0) * B(n))
           YV = ((c+1.0)*Bxsum1) + ((c-1.0)*B(n)*x(n)) - (D*x(1))
           ablu = x(1) - YV/((nc-1.0)*(nc-2.0) - (D*nc))
           bblu = (nc-1.0) * (x(1) - ablu)
c
         return
         end
c
c==================================================================
c                         END SUBROUTINE BLCGT2
c******************************************************************
```

```
      Subroutine HYPCDF
c*****************************************************************
c**                                                            **
c**       B E G I N   S U B R O U T I N E   H Y P C D F        **
c**                                                            **
c*****************************************************************
c
c  Ref:  Appendix A, Figure 6, Step 4.
c
c=================================================================
c
c  Purpose:  Given an ordered sample of size n, a specified
c            shape c, and the BLUEs of location a and scale b,
c            compute the hypothesized Pareto distribution
c            function P(i) for i = 1,2,...,n.
c
c=================================================================
c
c  Variables:
c                x = array containing n ordered Pareto deviates
c                n = sample size
c                c = shape parameter
c             ablu = BLUE of location a
c             bblu = BLUE of scale b
c                P = array containing n points of the
c                    hypothesized Pareto CDF
c
c=================================================================
c
c  Input:
c         x = array of n ordered Pareto deviates (from PARDEV)
c         c = shape parameter = .5(.5)4  (from MAIN DO Loop 90)
c         n = sample size = 5(5)30  (from MAIN DO Loop 80)
c      ablu = BLUE of location a (from BLCLE2 or BLCGT2)
c      bblu = BLUE of scale b (from BLCLE2 or BLCGT2)
c
c================================================================
c
c  Calculate:
c
c     P(i) = 1 - [1 / [1 + (x(i) - ablu)/bblu] ]**c     (eqn 40)
c
c================================================================
c
c  Output:  P = array of n points of the hypothesized CDF
c
c================================================================
c
c  Declare Variables:
```

```
c
      common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
     1        Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
     1        KScrit,ADcrit,CVcrit,Y
      integer n
      real x(30),ablu,bblu,B(30),D,KS(5000,6,8),AD(5000,6,8),
     1     CVM(5000,6,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,P(30),
     1     KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),r(30),
     1     Y(5002)
      double precision dseed
c
      do 10 i = 1,n
         P(i) = 1.0 - (1.0 + (x(i) - ablu)/bblu) ** (-c)
   10 continue
c
      return
      end
c
c=====================================================================
c                         END SUBROUTINE HYPCDF
c*********************************************************************
```

```
      Subroutine TESTAT
c********************************************************************
c**                                                                **
c**      B E G I N   S U B R O U T I N E   T E S T A T             **
c**                                                                **
c********************************************************************
c
c  Ref:  Appendix A, Figure 6, Step 5.
c
c====================================================================
c  Purpose:  Given a sample size n, and the hypothesized Pareto
c            distribution function P(i), compute values of the
c            test statistics of the modified K-S, A-D, and CVM
c            goodness-of-fit tests.
c====================================================================
c  Variables:
c                n = sample size
c             nshp = shape parameter counter (8 values, 1-8)
c             nsiz = sample size counter (6 values, 1-6)
c               it = iteration counter (1-5000)
c                P = array of n values of the hypothesized Pareto CDF
c          ----------------------------------------------------------
c               DP = positive differences between EDF and CDF points
c               DM = negative differences between EDF and CDF points
c            DPLUS = maximum positive difference (largest DP value)
c           DMINUS = maximum negative difference (largest DM value)
c               KS = values of the modified K-S test statistic
c          ----------------------------------------------------------
c               AL = value used to calculate the A-D test statistic
c               AM = value used to calculate the A-D test statistic
c               AN = AL + AM
c              AAA = values to be summed for A-D test statistic
c             SAAA = sum of AAA values
c               AD = values of the modified A-D test statistic
c          ----------------------------------------------------------
c              ACV = squared quantities in the C-VM formula
c             SACV = sum of the ACV values
c              CVM = values of the modified C-VM test statistic
c====================================================================
c  Input:
c             n = sample size = 5(5)30  (from MAIN DO Loop 80)
c             P = array of n values of hypothesized CDF (from HYPCDF)
c            it = iteration counter (from MAIN Do Loop 60)
c         nsiz = sample size counter (from MAIN DO Loop 80)
c         nshp = shape parameter counter (from MAIN DO Loop 90)
c====================================================================
c  Calculations for K-S test statistic (eqns 41 & 42):
c
c              DP(i) = ABS[ (i/n) - P(i) ]
c              DM(i) = ABS[ P(i) - (i-1)/n ]
c
c              DPLUS = max [ DP(i) ]  for i=1,2,...,n
```

```
c              DMINUS = max [ DM(i) ]  for i=1,2,...,n
c
c                 KS = max  (DPLUS,DMINUS)
c
c-----------------------------------------------------------------
c
c  Calculations for A-D test statistic (eqn 43):
c
c               AL(j) = ln (P(j))
c               AM(j) = ln (1 - P(n+1-j))
c               AN(j) = AL(j) + AM(j)
c
c              AAA(j) = (2*j - 1) * AN(j)
c                SAAA = AAA(1) + AAA(2) + ... + AAA(n)
c
c                  AD = -n - (1/n) * SAAA
c
c-----------------------------------------------------------------
c
c  Calculations for C-VM test statistic (eqn 44):
c
c              ACV(k) = [ P(k) - (2*k - 1)/(2*n) ]**2
c                SACV = ACV(1) + ACV(2) + ... + ACV(n)
c
c                 CVM = (1/(12*n)) + SACV
c
c
c=================================================================
c
c  Declare Variables:
c
      common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
   1         Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
   1         KScrit,ADcrit,CVcrit,Y
      integer n,nsiz,nshp,it
      real x(30),ablu,bblu,B(30),D,KS(5000,6,8),AD(5000,6,8),
   1      CVM(5000,6,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,P(30),
   1      KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),r(30),
   1      DP(30),DM(30),DPLUS,DMINUS,AL(30),AM(30),
   1      AN(30),AAA(30),SAAA,ACV(30),SACV,Y(5002)
      double precision dseed
c
      DPLUS = 0
      DMINUS = 0
c
      do 5 ik = 1,30
         DP(ik) = 0
         DM(ik) = 0
  5   continue
c
c -----   Compute the K-S Test Statistic (eqns 41 & 42):  -----
c
```

```
          do 10 i = 1,n
              DP(i) = ABS( (i/real(n)) - P(i) )
              DM(i) = ABS( P(i) - (i-1)/real(n) )
   10     continue
c
          DPLUS =  MAX( DP(1),DP(2),DP(3),DP(4),DP(5),DP(6),DP(7),
     1               DP(8),DP(9),DP(10),DP(11),DP(12),DP(13),DP(14),
     1               DP(15),DP(16),DP(17),DP(18),DP(19),DP(20),
     1               DP(21),DP(22),DP(23),DP(24),DP(25),DP(26),
     1               DP(27),DP(28),DP(29),DP(30) )
c
          DMINUS = MAX( DM(1),DM(2),DM(3),DM(4),DM(5),DM(6),DM(7),
     1               DM(8),DM(9),DM(10),DM(11),DM(12),DM(13),DM(14),
     1               DM(15),DM(16),DM(17),DM(18),DM(19),DM(20),
     1               DM(21),DM(22),DM(23),DM(24),DM(25),DM(26),
     1               DM(27),DM(28),DM(29),DM(30) )
c
          KS(it,nsiz,nshp) = MAX(DPLUS,DMINUS)
c
c -------   Compute the A-D Test Statistic (eqn 43):   -------
c
          SAAA = 0
c
          do 20 j = 1,n
c
              AL(j) = log (P(j))
              AM(j) = log (1.0 - P(n+1-j))
              AN(j) = AL(j) + AM(j)
              AAA(j) = (2.0*j - 1.0) * AN(j)
              SAAA = SAAA + AAA(j)
c
   20     continue
c
          AD(it,nsiz,nshp) = -n - (1.0/real(n)) * SAAA
c
c -------   Compute the C-VM Test Statistic (eqn 44):   -------
c
          SACV = 0
c
          do 30 k = 1,n
              ACV(k) = ( P(k) - (2.0*k - 1.0)/(2.0*real(n)) )**2
              SACV = SACV + ACV(k)
   30     continue
c
          CVM(it,nsiz,nshp) = SACV + (1.0/(12.0*real(n)))
c
          return
          end
c
c=================================================================
c                        END SUBROUTINE TESTAT
c*****************************************************************
```

```
      Subroutine CRTVAL
c*******************************************************************
c**                                                               **
c**      B E G I N    S U B R O U T I N E    C R T V A L         **
c**                                                               **
c*******************************************************************
c
c  Ref: Appendix A, Figure 6, Step 7.
c
c===================================================================
c
c  Purpose:
c
c       Given a set of 5000 values of test statistics from the
c       modified Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D),
c       or Cramer-von Mises (C-VM) test, select critical values
c       by using median ranks plotting positions to compute
c       specified percentile levels.
c
c===================================================================
c
c  Variables:
c                 c = shape parameter
c                 n = sample size
c               pct = percentile value
c              nshp = shape parameter counter (1: c=.5; 2: c=1.0;
c                     3: c=1.5; 4: c=2.0; 5: c=2.5; 6: c=3.0;
c                     7: c=3.5; 8: c=4.0)
c              nsiz = sample size counter (1: n=5 or 6; 2: n=10;
c                     3: n=15; 4: n=20; 5: n=25; 6: n=30)
c              npct = percentile counter (0: pct=0; 1: pct=.80;
c                     2: pct=.85; 3: pct=.9; 4: pct=.95; 5: pct=.99)
c               nst = total number of statistics used
c                it = iteration counter (5000 repetitions required)
c                KS = 3D array of 5000 modified K-S test statistics
c               KS1 = 1D array of 5000 K-S test statistics
c               CVM = 3D array of 5000 modified C-VM test statistics
c               CV1 = 1D array of 5000 C-VM test statistics
c                AD = 3D array of 5000 modified A-D test statistics
c               AD1 = 1D array of 5000 A-D statistics
c              STAT = 1D array of test stats (KS, AD, or CVM)
c            KScrit = array of critical values for the K-S test
c           CVMcrit = array of critical values for the C-VM test
c            ADcrit = array of critical values for the A-D test
c              CRIT = either the KS, AD, or CVM critical value array
c                 Y = array containing 5002 plotting positions
c              slpm = array of slopes used to find critical values
c                bi = array of intercepts used to find critical vals
c
c==================================================================
c
c  Input:
```

```
c                  Y = array of plotting positions (MAIN DO Loop 10)
c                  c = shape parameter (from MAIN DO Loop 90)
c                  n = sample size (from MAIN DO Loop 80)
c               nshp = shape parameter counter (from MAIN DO Loop 90)
c               nsiz = sample size counter (from MAIN DO Loop 80)
c               npct = percentile counter (from MAIN DO Loop 70)
c                nst = number of test statistics used (from MAIN Prog)
c                 KS = array of 5000 K-S test statistics (from TESTAT)
c                CVM = array of 5000 C-VM test stats (from TESTAT)
c                 AD = array of 5000 A-D test statistics (from TESTAT)
c
c==================================================================
c
c  IMSL Subroutine:  VSRTA - orders the test statistic values
c
c==================================================================
c
c  Calculate Endpoints of Test Statistics (Eqns 52 - 57):
c
c              slpm(0) = ( Y(2) - Y(1) ) / ( STAT(2) - STAT(1) )
c                bi(0) = Y(1) - slpm(0) * STAT(1)
c              STAT(0) = max ( 0, - bi(0)/slpm(0) )
c
c              slpm(6) = (Y(5000) - Y(4999))/(STAT(5000) - STAT(4999))
c                bi(6) = Y(4999) - slpm(6) * STAT(4999)
c              STAT(6) = (1.0 - bi(6)) / slpm(6)
c
c------------------------------------------------------------------
c
c  Calculate Critical Values (Eqns 58 - 60):
c
c        slpm(npct) = ( Y(j+1) - Y(j) ) / ( STAT(j+1) - STAT(j) )
c          bi(npct) = Y(j) - slpm(npct) * STAT(j)
c        CRIT(npct) = (pct - bi(npct)) / slpm(npct)
c
c==================================================================
c
c  Output:
c
c    KScrit - array of critical values for modified K-S test
c    ADcrit - array of critical values for modified A-D test
c    CVcrit - array of critical values for modified C-VM test
c
c==================================================================
c
c  Declare Variables:
c
        common dseed,x,n,c,nc,B,D,ablu,bblu,P,pct,Bsum1,Bxsum1,
     1         Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,npct,nst,
     1         KScrit,ADcrit,CVcrit,Y
        integer n,nsiz,nshp,it,npct,nst,ntest
        real x(30),ablu,bblu,B(30),D,KS(5000,6,8),AD(5000,6,8),
```

```
     1       CVM(5000,6,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,P(30),
     1       KScrit(6,8,5),ADcrit(6,8,5),CVcrit(6,8,5),r(30),
     1       Y(5002),STAT(5002),CRIT(6,8,7),slpm(7),bi(7),pct,
     1       KS1(5000),CV1(5000),AD1(5000)
        double precision dseed
c
        if (npct .eq. 1) pct = .80
        if (npct .eq. 2) pct = .85
        if (npct .eq. 3) pct = .90
        if (npct .eq. 4) pct = .95
        if (npct .eq. 5) pct = .99
c
c     ** Store the 3 Sets of 5000 Test Stats into 1D Arrays: **
c
       do 16 ncnt = 1,nst
          KS1(ncnt) = KS(ncnt,nsiz,nshp)
          AD1(ncnt) = AD(ncnt,nsiz,nshp)
          CV1(ncnt) = CVM(ncnt,nsiz,nshp)
   16  continue
c
c     ** Use IMSL Subroutine to Order the Test Statistics: **
c
          Call VSRTA(KS1,nst)
c               print*,'ORDERED KS STATISTICS FROM CRTVAL:'
c               print*,'n=',n,' c=',c
c               do 2 jks = 1,nst
c                  print*,'KS STAT =',KS1(jks)
c 2             continue
c
          Call VSRTA(AD1,nst)
c               print*,'ORDERED AD STATISTICS FROM CRTVAL:'
c               print*,'n=',n,' c=',c
c               do 4 jad = 1,nst
c                  print*,'AD STAT =',AD1(jad)
c 4             continue
c
          Call VSRTA(CV1,nst)
c               print*,'ORDERED CVM STATISTICS FROM CRTVAL:'
c               print*,'n=',n,' c=',c
c               do 6 jcv = 1,nst
c                  print*,'CV STAT =',CV1(jcv)
c 6             continue
c
c --- Begin DO Loop 20 to Rotate Through KS, AD, and CVM ---
c
      do 20 ntest = 1,3
c
c     ---   Begin DO Loop 30 for 5000 Data Points   ---
c
         do 30 j = 1,nst
c
            if (ntest .eq. 1) then
```

```
                    STAT(j) = KS1(j)
                else if (ntest .eq. 2) then
                    STAT(j) = AD1(j)
                else if (ntest .eq. 3) then
                    STAT(j) = CV1(j)
                end if
c
   30       continue
c
c     ---   End DO Loop 30 for 5000 Data Points    ---
c
c     ** Extrapolate Left Endpoint of the Test Statistics: **
c
            if (STAT(1) .eq. STAT(2)) then
c
c               print*,'$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$'
c               print*,'TWO LEFT ENDPOINT STATS EQUAL'
c                  if (ntest .eq. 1) print*,'FOR KS TEST'
c                  if (ntest .eq. 2) print*,'FOR AD TEST'
c                  if (ntest .eq. 3) print*,'FOR CVM TEST'
c               print*,'n=',n,' c=',c,' pct=',pct
c               print*,'STAT(1)=',STAT(1)
c               print*,'STAT(2)=',STAT(2)
c               print*,'%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%'
c               print*,' '
c
                dif0 = STAT(3) - STAT(1)
                      if (dif0 .eq. 0.0) dif0 = .00001
                slpm(0) = (Y(3) - Y(1)) / dif0
            else
               dif0 = STAT(2) - STAT(1)
               slpm(0) = (Y(2) - Y(1)) / dif0
            end if
c
            bi(0) = Y(1) - slpm(0) * STAT(1)
            STAT(0) = max( 0.0, - bi(0)/slpm(0) )
c            print*,' '
c            print*,'========================================'
c              if (ntest .eq. 1)print*,'FOR KS TEST STATISTICS'
c              if (ntest .eq. 2)print*,'FOR AD TEST STATISTICS'
c              if (ntest .eq. 3)print*,'FOR CVM TEST STATISTICS'
c            print*,'LEFT ENDPT X(0000) =',STAT(0)
c            print*,'-----FIRST X(0001) =',STAT(1)
c            print*,'80PCT STAT X(4000) =',STAT(4000)
c            print*,'85PCT STAT X(4250) =',STAT(4250)
c            print*,'90PCT STAT X(4500) =',STAT(4500)
c            print*,'95PCT STAT X(4750) =',STAT(4750)
c            print*,'99PCT STAT X(4950) =',STAT(4950)
c            print*,'----- LAST X(5000) =',STAT(5000)
c
c     ** Extrapolate Right Endpoint of the Test Statistic: **
c
```

```
                  if (STAT(nst-1) .eq. STAT(nst)) then
c
c                    print*,'$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$'
c                    print*,'TWO RIGHT ENDPOINT STATS EQUAL:'
c                       if (ntest .eq. 1) print*,'FOR KS TEST'
c                       if (ntest .eq. 2) print*,'FOR AD TEST'
c                       if (ntest .eq. 3) print*,'FOR CVM TEST'
c                    print*,'n=',n,' c=',c,' pct=',pct
c                    print*,'STAT(4999)=',STAT(nst-1)
c                    print*,'STAT(5000)=',STAT(nst)
c                    print*,'%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%'
c                    print*,' '
c
                     dif6 = STAT(nst) - STAT(nst-2)
                           if (dif6 .eq. 0.0) dif6 = .00001
                     slpm(6) = (Y(nst)-Y(nst-2)) / dif6
                  else
                     dif6 = STAT(nst) - STAT(nst-1)
                     slpm(6) = (Y(nst)-Y(nst-1)) / dif6
                  end if
c
                  bi(6) = Y(nst-1) - slpm(6)*STAT(nst-1)
                  STAT(nst+1) = (1.0 - bi(6)) / slpm(6)
c                  print*,'RGHT ENDPT X(5001) =',STAT(nst+1)
c
c              ** Interpolate Critical Values Between Test Stats:  **
c
c                 -- Begin DO Loop 50 to Find Max Y(k) < pct: --
c
                     do 50 kj = 1,nst
                        k = nst+1 - kj
c
                        if (Y(k) .le. pct) then
c
                  if (STAT(k) .eq. STAT(k+1)) then
c
c                    print*,'$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$'
c                    print*,'TWO ADJACENT STATS EQUAL:'
c                       if (ntest .eq. 1) print*,'FOR KS TEST'
c                       if (ntest .eq. 2) print*,'FOR AD TEST'
c                       if (ntest .eq. 3) print*,'FOR CVM TEST'
c                    print*,'n=',n,' c=',c,' pct=',pct
c                    print*,'STAT(k)=',STAT(k)
c                    print*,'STAT(k+1)=',STAT(k+1)
c                    print*,'%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%'
c                    print*,' '
c
                     dif = STAT(k+1) - STAT(k-1)
                          if (dif .eq. 0.0) dif = .00001
                     slpm(npct) = (Y(k+1)-Y(k-1)) / dif
                  else
                     dif = STAT(k+1) - STAT(k)
```

```
                  slpm(npct) = (Y(k+1)-Y(k)) / dif
               end if
c
                  bi(npct) = Y(k) - slpm(npct) * STAT(k)
                  CRIT(nsiz,nshp,npct)
     1                          = (pct-bi(npct))/slpm(npct)
                     GOTO 75
c
                    end if
c
  50              continue
c
c               -- End DO Loop 50 Upon Finding Crit Val --
c
c        **  Associate the Critical Values with Test Type:  **
c
  75            if (ntest .eq. 1) then
                KScrit(nsiz,nshp,npct) =  CRIT(nsiz,nshp,npct)
c                print*,'n=',n,'  **  c=',c,' pct=',pct
c                print*,'CRTVAL KS Crit Val =',KScrit(nsiz,nshp,npct)
             else if (ntest .eq. 2) then
                ADcrit(nsiz,nshp,npct) = CRIT(nsiz,nshp,npct)
c                print*,'CRTVAL AD Crit Val =',ADcrit(nsiz,nshp,npct)
             else if (ntest .eq. 3) then
                CVcrit(nsiz,nshp,npct) =  CRIT(nsiz,nshp,npct)
c                print*,'CRTVAL CV Crit Val =',CVcrit(nsiz,nshp,npct)
c                print*,' '
             end if
c
  20  continue
c
c --- End DO Loop 20 After Rotating Through KS, AD, and CVM ---
c
      return
      end
c
c=================================================================
c                    END SUBROUTINE CRTVAL
c*****************************************************************
```

APPENDIX B

Computer Program and Subroutines

for Determining Power Values

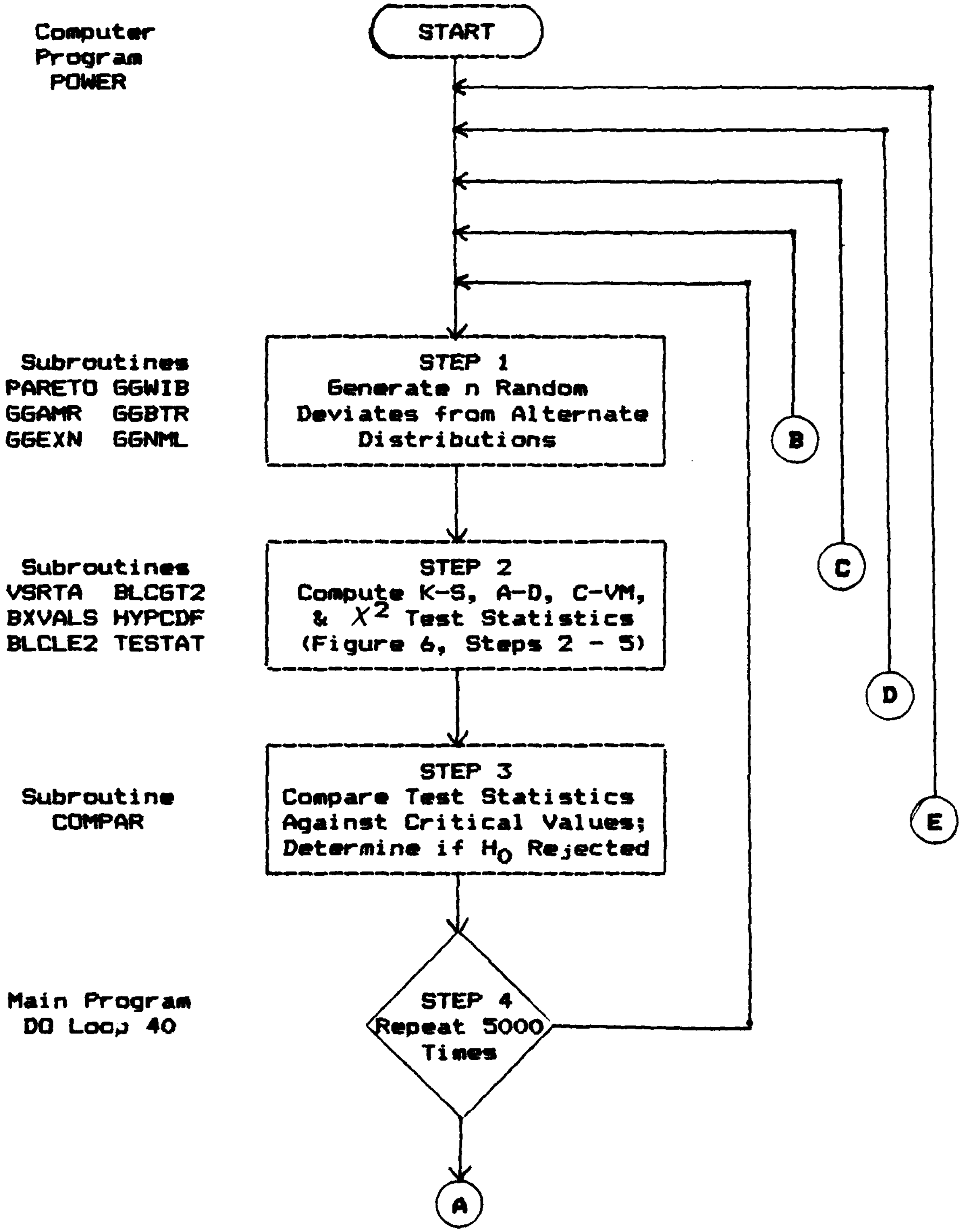


Fig 7. Procedure for Determining Power Values

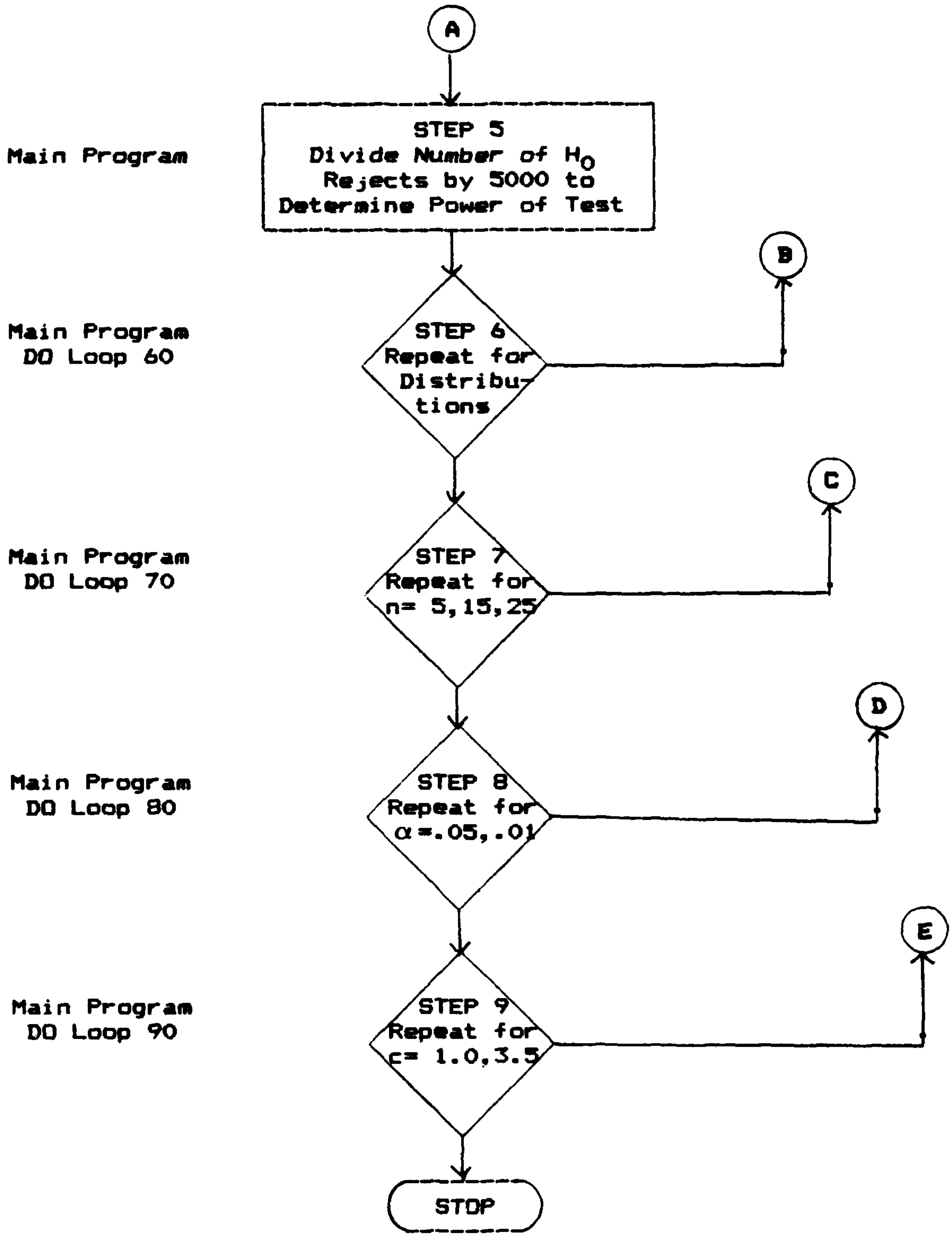


Fig 7 (Continued). Procedure for Determining Power Values

```
c***** Classroom Support Computer (CSC) - VAX 11/785 - VMS 4.1 ****
c
c******    POWER PROGRAM FOR PARETO GOODNESS-OF-FIT TESTS    ******
c
c*****************************************************************
c*****************************************************************
c**                                                             **
c**     B E G I N    P O W E R    M A I N    P R O G R A M    **
c**                                                             **
c*****************************************************************
c
c  Ref: Appendix B, Figure 7.
c
c=================================================================
c
c  Purpose:  Test the null hypothesis that a set of sample data
c     follows the Pareto distribution with hypothesized shape c
c     against the alternate hypothesis that the data follow some
c     other distribution.  The goals are to:
c
c     1.  Compare powers of the modified Kolmogorov-Smirnov (K-S),
c     Anderson-Darling (A-D), and Cramer-von Mises (C-VM) tests
c     against the Chi-Square test to determine which test can
c     best detect a false Pareto distribution hypothesis.
c
c     2.  When the Pareto null hypothesis is true, confirm that
c     the hypothesis rejection rates under the modified K-S, A-D,
c     and C-VM statistics are low enough to satisfy a claimed level
c     of significance.
c
c     3.  Provide extensive commentary to assist novice programmers
c     to conduct similar power studies in statistical analysis.
c     Diagnostic print statements have been retained as commentary
c     to contribute to this goal.
c
c=================================================================
c
c  Variables:
c          dseed = random number seed
c          alpha = level of significance (.01 or .05 used here)
c              n = sample size
c              c = null-hypothesis Pareto shape parameter
c           nshp = null-hyp Pareto shape counter (1:c=1.0, 2:c=3.5)
c           nalf = significance level counter (1: α =.05, 2: α =.01)
c           nsiz = sample size counter (1:n=5, 2:n=15, 3:n=25)
c           nalt = alternative distribution counter (8 in all)
c           nrep = number of repetitions to be used
c             it = iteration counter (5000 repetitions required)
c             KS = array of values of modified K-S test statistic
c            CVM = array of values of modified C-VM test statistic
c             AD = array of values of modified A-D test statistic
c             X2 = array of values of Chi-square test statistic
```

```
c             nrKS = number of hypothesis rejects under the K-S test
c             nrAD = number of hypothesis rejects under the A-D test
c             nrCV = number of hypothesis rejects under the CVM test
c             nrX2 = number of hypothesis rejects under Chi-square
c
c===================================================================
c
c  Input:
c      nrep = number of repetitions (input at computer terminal)
c     dseed = random number seed (input at computer terminal)
c
c===================================================================
c
c  Subroutines:
c
c    PARETO - Generates n random Pareto deviates
c    BXVALS - Calculates B values and summations of B and Bx
c    BLCLE2 - Finds BLUEs for location and scale when c <= 2
c    BLCGT2 - Finds BLUEs for location and scale when c > 2
c    HYPCDF - Computes the Hypothesized Pareto CDF
c    TESTAT - Calculates the K-S, A-D, and C-VM test statistics
c    COMPAR - Compares test stats vs. crit vals and counts rejects
c
c------------------------------------------------------------------
c
c  IMSL Subroutines:
c
c     GGWIB - Generates random Weibull deviates
c     GGAMR - Generates random Gamma deviates
c     GGBTR - Generates random Beta deviates
c     GGEXN - Generates random Exponential Deviates
c     GGNML - Generates random Normal Deviates
c     VSRTA - Arranges data in ascending order
c
c==================================================================
c
c  Output:
c
c     KSpwr(nshp,nalf,nsiz,nalt) = power values for K-S test
c     ADpwr(nshp,nalf,nsiz,nalt) = power values for A-D test
c     CVpwr(nshp,nalf,nsiz,nalt) = power values for C-VM test
c     X2pwr(nshp,nalf,nsiz,nalt) = power values for Chi-square
c
c=================================================================
c
c  Declare Variables:
c
        common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1          Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,nrep,
     1          nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
        integer n,nsiz,nshp,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1          nrCV(2,2,3,8),nrX2(2,2,3,8)
```

```
      real    x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1        CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1        P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1        CVpwr(2,2,3,8),X2crit(2,2,3),X2(2,2,3,8),
     1        X2pwr(2,2,3,8)
      character test(4)*3,altcdf(8)*12
      double precision dseed
c
            test(1) = 'K-S'
            test(2) = 'A-D'
            test(3) = 'CVM'
            test(4) = 'CHI'
c
            altcdf(1) = 'Pareto c=1.0'
            altcdf(2) = 'Pareto c=3.5'
            altcdf(3) = 'Pareto c=2.0'
            altcdf(4) = 'Weibull'
            altcdf(5) = 'Gamma'
            altcdf(6) = 'Beta'
            altcdf(7) = 'Exponential'
            altcdf(8) = 'Normal'
c
c    ** Open Output File to Store Computed Power Values: **
            open (unit=7,file='X2ALL',status='new')
c
c    **  Number of Repetitions to be Used on Each Run:  **
            print*,'The Monte Carlo power analysis will require'
            print*,'     5000 repetitions.'
            print*,'Enter the number to be used for this run:'
            read*,nrep
c
    print*,'Enter random number seed or "1." for default:'
    read*,dseed
         if (dseed .eq. 1.) dseed = 123457.d00
    print*,' '
    print*,'STANDBY . . . COMPUTATIONS IN PROGRESS'
c
c ---  Begin DO Loop 90 for Null-Hypothesis Pareto Shape c  ---
c
    do 90 nshp = 1,2
c
       if (nshp .eq. 1) then
          c = 1.0
          write(7,51)
          write(7,56)
          write(7,58)
          write(7,62)
       else if (nshp .eq. 2) then
          c = 3.5
          write(7,52)
          write(7,56)
          write(7,59)
```

```
            write(7,62)
         end if
c
c     --- Begin DO Loop 80 for Alpha Significance Levels ---
c
         do 80 nalf =1,2
c
            if (nalf .eq. 1) then
               alpha = .05
               write(7,64)
            else if (nalf .eq. 2) then
               alpha = .01
               write(7,66)
            end if
c
           write(7,54)
           write(7,74)
           write(7,68)
           write(7,72)
           write(7,76)
           write(7,72)
c
           nsiz = 0
c
c        print*,'========================='
c        print*,'Numbers of Rejects After do 80/Before do 70'
c        print*,'c =',c,'alpha =',alpha,'n=',n,'CDF: ',altcdf(nalt)
c        print*,'KS Rejects = ',nrKS(nshp,nalf,nsiz,nalt)
c        print*,'AD Rejects = ',nrAD(nshp,nalf,nsiz,nalt)
c        print*,'CV Rejects = ',nrCV(nshp,nalf,nsiz,nalt)
c        print*,'========================='
c
c          ---    Begin DO Loop 70 for Sample Sizes    ---
c
            do 70 n = 5,25,10
c
               nsiz = nsiz + 1
c
               nc = n * c
c
c             -- Begin DO Loop 60 for Alternate CDFs  --
c
               do 60 nalt = 1,8
c
c
               nrKS(nshp,nalf,nsiz,nalt) = 0
               nrAD(nshp,nalf,nsiz,nalt) = 0
               nrCV(nshp,nalf,nsiz,nalt) = 0
               nrX2(nshp,nalf,nsiz,nalt) = 0
c
c              -- Begin DO Loop 40 for Repetitions --
c
```

```
                 do 40 it = 1,nrep
c
c              ** Perform Step 1 of Figure 7: **
c
               if (nalt .eq. 1) call PARETO
               if (nalt .eq. 2) call PARETO
               if (nalt .eq. 3) call PARETO
               if (nalt .eq. 4) call GGWIB(dseed,3.5,n,x)
               if (nalt .eq. 5) call GGAMR(dseed,2.,n,1,x)
               if (nalt .eq. 6) call GGBTR(dseed,2.,3.,n,x)
               if (nalt .eq. 7) call GGEXN(dseed,2.,n,x)
               if (nalt .eq. 8) call GGNML(dseed,n,x)
c
c              ** Perform Step 2 of Figure 7: **
c
                   call VSRTA(x,n)
                   call BXVALS
c
                   if (c .eq. 1.0) call BLCLE2
                   if (c .eq. 3.5) call BLCGT2
c
                   call HYPCDF
                   call TESTAT
c
c              ** Perform Step 3 of Figure 7: **
c
                   call COMPAR
c
   40              continue
c
c              --  End DO Loop 40 for Repetitions  --
c              ** Completes Step 4 of Figure 7 **
c
c              **  Perform Step 5 of Figure 7: **
c
c           print*,'==========================='
c           print*,'Numbers of Rejects Prior to Power Calculation'
c           print*,'c =',c,'alpha =',alpha,'n=',n,'nalt=',nalt
c           print*,'KS Rejects = ',nrKS(nshp,nalf,nsiz,nalt)
c           print*,'AD Rejects = ',nrAD(nshp,nalf,nsiz,nalt)
c           print*,'CV Rejects = ',nrCV(nshp,nalf,nsiz,nalt)
c           print*,'X2 Rejects = ',nrX2(nshp,nalf,nsiz,nalt)
c           print*,'==========================='
c
                    KSpwr(nshp,nalf,nsiz,nalt) =
     1         nrKS(nshp,nalf,nsiz,nalt)/real(nrep)
c
                    ADpwr(nshp,nalf,nsiz,nalt) =
     1         nrAD(nshp,nalf,nsiz,nalt)/real(nrep)
c
                    CVpwr(nshp,nalf,nsiz,nalt) =
     1         nrCV(nshp,nalf,nsiz,nalt)/real(nrep)
```

```
c
           X2pwr(nshp,nalf,nsiz,nalt) =
     1  nrX2(nshp,nalf,nsiz,nalt)/real(nrep)
c
          print*,'*****************************************'
          print*,' POWER VALUES FROM MAIN PROGRAM'
          print*,' Null-hyp c =',c,'alpha =',alpha
          print*,' n=',n,'    Alternate CDF: ',altcdf(nalt)
c          print*,'=========================='
c          print*,' KS Rejects = ',nrKS(nshp,nalf,nsiz,nalt)
c          print*,' AD Rejects = ',nrAD(nshp,nalf,nsiz,nalt)
c          print*,' CV Rejects = ',nrCV(nshp,nalf,nsiz,nalt)
c          print*,' X2 Rejects = ',nrX2(nshp,nalf,nsiz,nalt)
c          print*,'=========================='
          print*,' KS Power =',KSpwr(nshp,nalf,nsiz,nalt)
          print*,' AD Power =',ADpwr(nshp,nalf,nsiz,nalt)
          print*,' CV Power =',CVpwr(nshp,nalf,nsiz,nalt)
          print*,' X2 Power =',X2pwr(nshp,nalf,nsiz,nalt)
          print*,'*****************************************'
          print*,' '
c
   60    continue
c
c       -- End DO Loop 60 for Alternate CDFs --
c          ** Completes Step 6 of Figure 7 **
c
c       --   Write Power Results to File    --
c
        write(7,110),n,test(1),KSpwr(nshp,nalf,nsiz,1),
     1   KSpwr(nshp,nalf,nsiz,2),KSpwr(nshp,nalf,nsiz,3),
     1   KSpwr(nshp,nalf,nsiz,4),KSpwr(nshp,nalf,nsiz,5),
     1   KSpwr(nshp,nalf,nsiz,6),KSpwr(nshp,nalf,nsiz,7),
     1   KSpwr(nshp,nalf,nsiz,8)
c
        write(7,110),n,test(2),ADpwr(nshp,nalf,nsiz,1),
     1   ADpwr(nshp,nalf,nsiz,2),ADpwr(nshp,nalf,nsiz,3),
     1   ADpwr(nshp,nalf,nsiz,4),ADpwr(nshp,nalf,nsiz,5),
     1   ADpwr(nshp,nalf,nsiz,6),ADpwr(nshp,nalf,nsiz,7),
     1   ADpwr(nshp,nalf,nsiz,8)
c
        write(7,110),n,test(3),CVpwr(nshp,nalf,nsiz,1),
     1   CVpwr(nshp,nalf,nsiz,2),CVpwr(nshp,nalf,nsiz,3),
     1   CVpwr(nshp,nalf,nsiz,4),CVpwr(nshp,nalf,nsiz,5),
     1   CVpwr(nshp,nalf,nsiz,6),CVpwr(nshp,nalf,nsiz,7),
     1   CVpwr(nshp,nalf,nsiz,8)
c
        write(7,110),n,test(4),X2pwr(nshp,nalf,nsiz,1),
     1   X2pwr(nshp,nalf,nsiz,2),X2pwr(nshp,nalf,nsiz,3),
     1   X2pwr(nshp,nalf,nsiz,4),X2pwr(nshp,nalf,nsiz,5),
     1   X2pwr(nshp,nalf,nsiz,6),X2pwr(nshp,nalf,nsiz,7),
     1   X2pwr(nshp,nalf,nsiz,8)
c
```

```
                   write(7,72)
c
  70               continue
c
c            ---    End DO Loop 70 for Sample Sizes   ---
c                  ** Completes Step 7 of Figure 7 **
c
  80             continue
c
c   ---  End DO Loop 80 for Alpha Significance Levels ---
c                  ** Completes Step 8 of Figure 7 **
c
                write(7,74)
c
  90   continue
c
c --- End DO Loop 90 for Null-Hypothesis Pareto Shape Parameter ---
c
c*******************************************************************
c
c      Specify Format for Hardcopy Output Data and Headers:
c
  51  format('1',36X,'Table XVII')
  52  format('1',35X,'Table XVIII')
  54  format(' ')
  56  format('0',22X,'POWER TEST FOR THE PARETO DISTRIBUTION')
  58  format(22X,'Ho: Pareto Distribution at Shape c = 1.0')
  59  format(22X,'Ho: Pareto Distribution at Shape c = 3.5')
  62  format(22X,'Ha: The data follow another distribution')
  64  format('0',28X,'Level of Significance   = .05')
  66  format('0',28X,'Level of Significance   = .01')
  68  format(35X,'Alternate Distributions')
  72  format(80('-'))
  74  format(80('='))
  76  format(2X,' n',3X,'Test',4X,'Par.1',3X,'Par.2',3X,'Par.3',3X,
     1       'Weibl',3X,'Gamma',3X,'Beta',4X,
     1       'Expon',3X,'Norml')
 110  format(' ',I3,A7,F9.3,7F8.3)
c
      close(7)
c
      end
c
c==================================================================
c                            END MAIN PROGRAM
c*******************************************************************
```

```
      Subroutine PARETO
c*******************************************************************
c**                                                               **
c**      B E G I N   S U B R O U T I N E   P A R E T O            **
c**                                                               **
c*******************************************************************
c
c  Ref:  Appendix B, Fig 7, Step 1.
c
c==================================================================
c
c  Purpose:  For a specified sample size n, generate n random
c            deviates from a Pareto distribution with parameters of
c            location, scale, and shape set to specified positive
c            values.
c
c==================================================================
c
c  Variables:
c                    r = array containing n random numbers
c                   ac = actual shape parameter of Pareto deviates
c                    x = array containing n Pareto deviates
c                    n = sample size
c                dseed = random number seed
c
c==================================================================
c
c  Input:      dseed = random number seed  (from MAIN program)
c                  n = sample size = 5,15, or 25  (MAIN DO Loop 70)
c               nalt = alternate CDF counter (MAIN DO Loop 60)
c
c==================================================================
c
c  IMSL Subroutines:
c
c   GGUBFS - generates random numbers distrib uniformly on (0,1)
c   VSRTA  - arranges a set of numbers in ascending order
c
c==================================================================
c
c  Calculate:
c
c      x(j) = (1/r(j)) ** (1/ac)   for j = 1,2,....,n  (from eqn 48)
c
c   x(a',b') = b' * ( ( x(a,b) - a ) / b) + a'  (from eqn 50)
c
c==================================================================
c
c  Output:      x = array of n random Pareto deviates
c
c==================================================================
c
```

```
c  Declare Variables:
c
        common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1          Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nsho,nrep,
     1          nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
        integer n,nsiz,nsho,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1          nrCV(2,2,3,8)
        real    x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1          CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1          P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1          CVpwr(2,2,3,8),ac
        double precision dseed
c
          if (nalt .eq. 1) ac = 1.0
          if (nalt .eq. 2) ac = 3.5
          if (nalt .eq. 3) ac = 2.0
c
c--- Begin DO Loop 10 to Generate n Random Pareto Deviates ---
c
        do 10 j = 1,n
c
c            Use IMSL subroutine to generate random numbers:
                r(j) = GGUBFS(dseed)
c
c            Use eqn 48 to transform them to Pareto deviates
c            with location a = 1 and scale b = 1:
                x(j) = (1.0/r(j))**(1.0/ac)
c
c            Use eqn 50 to transform to Pareto deviates with
c            a = 2, b = 3 for the second alternate CDF:
                if (nalt .eq. 2) x(j) = 3. * x(j) - 1.
c
c            Use eqn 50 to transform to Pareto deviates with
c            a = 10, b = 5 for the third alternate CDF:
                if (nalt .eq. 3) x(j) = 5. * x(j) + 5.
c
   10   continue
c
c---   End DO Loop 10 after Generating n Random Deviates  ---
c
        return
        end
c
c==================================================================
c                       END SUBROUTINE PARETO
c******************************************************************
```

```
      Subroutine BXVALS
c*********************************************************************
c**                                                                 **
c**      B E G I N    S U B R O U T I N E    B X V A L S            **
c**                                                                 **
c*********************************************************************
c
c  Ref:  Appendix B, Fig. 7, Step 2.
c
c=====================================================================
c
c  Purpose:  For a given sample size n, calculate the B values
c            used to find the BLUEs of location and scale.  Also
c            find the sum of the first n-1 values of B(i).  Then,
c            compute the three values equal to the sums of the
c            first n-1, the first n-2, and (for hypothesized
c            c = .5, 1, or 2) the first n -2/c values of B(i)x(i).
c
c=====================================================================
c
c  Variables:   c = null-hypothesis shape parameter
c               n = sample size
c               x = array containing n ordered deviates
c                   from an alternate distribution
c               B = array containing n values of B
c           Bsum1 = sum of B(i) values for i = 1,2,....,(n-1)
c          Bxsum1 = sum of B(i)x(i) for i = 1,2,....,(n-1)
c          Bxsum2 = sum of B(i)x(i) for i = 1,2,....,(n-2)
c          Bxsm2c = sum of B(i)x(i) for i = 1,2,...,(n-2/c)
c
c=====================================================================
c
c  Input:     c = null-hyp shape parameter   (from MAIN DO Loop 90)
c             n = sample size = 5, 15, or 25 (from MAIN DO Loop 70)
c            nc = n*c    (from MAIN program)
c             x = ordered deviates of alternate CDF  MAIN)
c
c=====================================================================
c
c  Calculate:
c
c          B(i) = [1 - 2/c(n-i+1)] * B(i-1)       (eqn 29)
c
c         Bsum1 = B(1) + B(2) + ... + B(n-1)
c
c        Bxsum1 = B(1)*x(1) + ... + B(n-1)*x(n-1)
c
c        Bxsum2 = B(1)*x(1) + ... + B(n-2)*x(n-2)
c
c        Bxsm2c = B(1)*x(1) + ... + B(n-2/c)*x(n-2/c)
c
c=====================================================================
```

```
c
c  Output:
c            B = array containing n values of B
c        Bsum1 = sum of first (n-1) B values
c       Bxsum1 = sum of first (n-1) B*x values
c       Bxsum2 = sum of first (n-2) B*x values
c       Bxsm2c = sum of first (n-2/c) B*x (if 2/c is integer)
c
c======================================================================
c
c  Declare Variables:
c
      common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1        Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,nrep,
     1        nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
      integer n,nsiz,nshp,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1        nrCV(2,2,3,8)
      real    x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1        CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1        P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1        CVpwr(2,2,3,8)
      double precision dseed
c
c  Calculate the first B value (eqn 25):
c
      B(1) = 1.0 - 2.0/nc
c
c --- Begin DO Loop 10 to Find the 2nd thru nth B values ---
c
      do 10 j = 2,n
         B(j) = B(j-1) * (1.0 - (2.0/(c*(n-j+1))))
  10  continue
c
c     --- End DO Loop 10 ---
c
      Bsum1 = 0
c
c --- Begin DO Loop 20 to Sum the First n-1 Values of B  ---
c
      do 20 k=1,(n-1)
         Bsum1 = Bsum1 + B(k)
  20  continue
c
c     --- End DO Loop 20 ---
c
      Bxsum1 = 0
c
c --- Begin DO Loop 30 to Sum the First n-1 Values of Bx ---
c
      do 30 l=1,(n-1)
         Bxsum1 = Bxsum1 + (B(l)*x(l))
  30  continue
```

```
c
c      --- End DO Loop 30 ---
c
        Bxsum2 = Bxsum1 - (B(n-1)*x(n-1))
c
c --- Find Bxsm2c When 2/c is an Integer (c=.5, 1, or 2) ---
c
        Bxsm2c = 0
c
        if (c .eq. 1.0) then
           Bxsm2c = Bxsum2
        else if (c .eq. 2.0) then
           Bxsm2c = Bxsum1
        else if (c .eq. 0.5) then
           Bxsm2c = Bxsum2 - (B(n-3)*x(n-3)) - (B(n-2)*x(n-2))
        end if
c
        return
        end
c
c=================================================================
c                      END SUBROUTINE BXVALS
c*****************************************************************
```

```
      Subroutine BLCLE2
c*******************************************************************
c**                                                              **
c**        B E G I N   S U B R O U T I N E   B L C L E 2         **
c**                                                              **
c*******************************************************************
c
c  Ref:  Appendix B, Figure 7, Step 2 (continued).
c
c===================================================================
c
c  Purpose:  Given an ordered sample of size n and null-hypothesis
c            c<=2, calculate the BLUEs of location a and scale b.
c
c===================================================================
c  Variables:
c                 x = array containing n ordered deviates from a CDF
c                 c = null-hypothesis Pareto shape parameter
c                 n = sample size
c                 B = array of B values used to calculate the BLUEs
c                nc = product of n and c
c             Coef1 = coefficient used to compute BLUE of location a
c             Coef2 = coefficient used to compute BLUE of location a
c             Coef3 = coefficient used to compute BLUE of scale b
c            Bxsum2 = sum of B(i)*x(i) terms for i = 1,....,n-2
c            Bxsm2c = sum of B(i)*x(i) terms for i = 1,....,n-2/c
c              ablu = BLUE of the location parameter a
c              bblu = BLUE of the scale parameter b
c                 U = value used to compute BLUEs when c = 1.5
c             Termi = terms used to compute U (i=1,2,3)
c===================================================================
c  Input:      x = array of n ordered deviates (from MAIN Program)
c              c = null-hyp shape = 1.0 (from MAIN DO Loop 90)
c              n = sample size = 5, 15, or 25 (from MAIN DO Loop 70)
c             nc = n*c (from MAIN program)
c              B = array containing n values of B (from BXVALS)
c         Bxsum2 = sum of first n-2 values of B (from BXVALS)
c         Bxsm2c = sum of first n-2/c values of B (from BXVALS)
c===================================================================
c  Calculate (if c = 0.5, 1.0, or 2.0):
c
c        Coef1 = [(c+1)*(c+2)] / [(nc-2)*(nc-c-2)]
c        Coef2 = (nc-2) / (c+2)
c
c         ablu = x(1) - Coef1 * [Bxsm2c - (Coef2*x(1))]  (eqn 34)
c         bblu = (nc-1) * [x(1) - ablu]    (eqn 35)
c-------------------------------------------------------------------
c  Calculate (if c = 1.5):
c
c        Term1 = (nc-2) * (nc-c-2)
c        Term2 = nc * (c-2) * B(n-1)
c        Term3 = (nc-1) * (c+2)
```

```
c          Coef3 = [(nc-1)/nc] * (nc-2-U)
c              U = (Term1 - Term2) / Term3    (eqn 39)
c
c           ablu = x(1) - bblu / (nc-1)    (eqn 37)
c           bblu = (1/U) * [(c+1)*(Bxsum2) + (2c-1)*B(n-1)*x(n-1)
c                  - Coef3 * x(1)]     (eqn 38)
c=====================================================================
c  Output:
c           ablu = BLUE of location parameter a
c           bblu = BLUE of scale parameter b
c=====================================================================
c
c  Declare Variables:
c
        common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1          Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,nrep,
     1          nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
        integer n,nsiz,nshp,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1          nrCV(2,2,3,8)
        real    x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1          CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1          P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1          CVpwr(2,2,3,8),Term1,Term2,Term3,Coef1,Coef2,
     1          Coef3,U
        double precision dseed
c
        if ((c.eq.0.5) .or. (c.eq.1.0) .or. (c.eq.2.0)) then
             Coef1 = ((c+1.0)*(c+2.0)) / ((nc-2.0)*(nc-c-2.0))
             Coef2 = (nc-2.0) / (c+2.0)
              ablu = x(1) - Coef1 * (Bxsm2c - (Coef2*x(1)))
              bblu = (nc-1.0) * (x(1) - ablu)
c
        else if (c .eq. 1.5) then
             Term1 = (nc-2.0) * (nc-c-2.0)
             Term2 = nc * (c-2.0) * B(n-1)
             Term3 = (nc-1.0) * (c+2.0)
                 U = (Term1 - Term2) / Term3
             Coef3 = ((nc-1.0)/nc) * (nc-2.0-U)
              bblu = (1.0/U) *( (c+1.0) * (Bxsum2)
     1                + (2.0*c-1.0)*B(n-1)*x(n-1) - Coef3 * x(1) )
              ablu = x(1) - (bblu / (nc-1.0))
c
        end if
c
        return
        end
c
c=====================================================================
c                        END SUBROUTINE BLCLE2
c*********************************************************************
```

```
      Subroutine BLCGT2
c*****************************************************************
c**                                                             **
c**     B E G I N   S U B R O U T I N E   B L C G T 2          **
c**                                                             **
c*****************************************************************
c
c  Ref:  Appendix B, Figure 7, Step 2 (continued).
c
c=================================================================
c
c  Purpose:  Given an ordered sample of size n and a Pareto null
c            hypothesis with shape c > 2, calculate the best
c            linear unbiased estimates (BLUEs) of location and
c            scale.
c
c=================================================================
c
c  Variables:  x = array containing n ordered deviates
c              c = null-hypothesis Pareto shape parameter
c              n = sample size
c             nc = product of n and c
c              B = array of B values used to calculate the BLUEs
c          Bsum1 = sum of B(i) terms for i = 1,...,n-1
c         Bxsum1 = sum of B(i)*x(i) terms for i = 1,...,n-1
c              D = value used to calculate the BLUEs
c             YV = value used to calculate the BLUEs
c           ablu = BLUE for location parameter a
c           bblu = BLUE for scale parameter b
c
c==================================================================
c
c  Input:      x = array of ordered deviates (from MAIN Program)
c              c = shape parameter = 3.5 (from MAIN DO Loop 90)
c              n = sample size = 5, 15, or 25 (MAIN DO Loop 70)
c             nc = n*c  (from MAIN Program)
c              B = array of B values (from BXVALS)
c          Bsum1 = sum of first (n-1) B values (from BXVALS)
c         Bxsum1 = sum of first n-1 B*x values (from BXVALS)
c
c==================================================================
c
c  Calculate:
c
c           D = [(c+1) * Bsum1] + [(c-1) * B(n)]          (eqn 21)
c
c        YV  = (c+1)*Bxsum1 + (c-1)*B(n)*x(n) - D*x(1)   (eqn 22)
c
c        ablu = x(1) - YV/[(nc-1)*(nc-2) - D*nc]          (eqn 17)
c
c        bblu = (nc-1) * [ x(1) - ablu ]                  (eqn 18)
c
```

```
c=================================================================
c
c  Output:     ablu = BLUE for location a
c              bblu = BLUE for scale b
c
c=================================================================
c
c  Declare Variables:
c
         common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1           Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,nrep,
     1           nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
         integer n,nsiz,nshp,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1           nrCV(2,2,3,8)
         real    x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1           CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1           P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1           CVpwr(2,2,3,8),YV
         double precision dseed
c
            D = ((c+1.0) * Bsum1) + ((c-1.0) * B(n))
            YV = ((c+1.0)*Bxsum1) + ((c-1.0)*B(n)*x(n)) - (D*x(1))
            ablu = x(1) - YV/((nc-1.0)*(nc-2.0) - (D*nc))
            bblu = (nc-1.0) * (x(1) - ablu)
c
         return
         end
c
c=================================================================
c                        END SUBROUTINE BLCGT2
c*****************************************************************
```

```
      Subroutine HYPCDF
c*****************************************************************
c**                                                             **
c**     B E G I N   S U B R O U T I N E   H Y P C D F           **
c**                                                             **
c*****************************************************************
c
c  Ref:  Appendix B, Figure 7, Step 2 (continued).
c
c=================================================================
c
c  Purpose:  Given an ordered sample of size n, a Pareto null-hyp
c            of shape c, and the BLUEs of location a and scale b,
c            compute the hypothesized Pareto distribution
c            function P(i) for i = 1,2,...,n.
c
c=================================================================
c
c  Variables:
c                 x = array containing n ordered deviates
c                 n = sample size
c                 c = null hypothesized Pareto shape parameter
c              ablu = BLUE of location a
c              bblu = BLUE of scale b
c                 P = array containing n points of the
c                     hypothesized Pareto CDF
c
c=================================================================
c
c  Input:
c            x = array of n ordered deviates (from MAIN Program)
c            c = null hyp shape = 1.0 or 3.5 (MAIN DO Loop 90)
c            n = sample size = 5, 15, or 25  (from MAIN DO Loop 70)
c         ablu = BLUE of location a (from BLCLE2 or BLCGT2)
c         bblu = BLUE of scale b (from BLCLE2 or BLCGT2)
c
c================================================================
c
c  Calculate:
c
c     P(i) = 1 - [1 + (x(i) - ablu)/bblu] ] ** (-c)     (eqn 40)
c
c================================================================
c
c  Output:  P = array of n points of the hypothesized CDF
c
c================================================================
c
c  Declare Variables:
c
      common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1        Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,nrep,
```

```
     1          nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
      integer n,nsiz,nshp,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1          nrCV(2,2,3,8)
      real      x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1          CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1          P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1          CVpwr(2,2,3,8)
      double precision dseed
c
      do 10 i = 1,n
         P(i) = 1.0 - (1.0 + (x(i) - ablu)/bblu) ** (-c)
   10 continue
c
      return
      end
c
c==================================================================
c                         END SUBROUTINE HYPCDF
c******************************************************************
```

```
      Subroutine TESTAT
c*****************************************************************
c**                                                             **
c**      B E G I N    S U B R O U T I N E    T E S T A T        **
c**                                                             **
c*****************************************************************
c
c  Ref:  Appendix B, Figure 7, Step 2.
c
c=================================================================
c
c  Purpose:  Given a sample size n, and the hypothesized Pareto
c            distribution function P(i), compute values of the
c            test statistics of the Chi-square and the modified
c            K-S, A-D, and CVM goodness-of-fit tests.
c
c=================================================================
c
c  Variables:
c              n = sample size
c           nshp = null-hyp shape counter (1: c=1.0, 2: c=3.5)
c           nalf = alpha level counter (1: α =.05, 2: α =.01)
c           nsiz = sample size counter (1: n=5, 2: n=15, 3: n=25)
c           nalt = alternate distribution counter
c              P = array of n values of the hypothesized Pareto CDF
c          ------------------------------------------------------
c             DP = positive differences between EDF and CDF points
c             DM = negative differences between EDF and CDF points
c          DPLUS = maximum positive difference (largest DP value)
c         DMINUS = maximum negative difference (largest DM value)
c             KS = values of the modified K-S test statistic
c          ------------------------------------------------------
c             AL = value used to calculate the A-D test statistic
c             AM = value used to calculate the A-D test statistic
c             AN = AL + AM
c            AAA = values to be summed for A-D test statistic
c           SAAA = sum of AAA values
c             AD = values of the modified A-D test statistic
c          ------------------------------------------------------
c            ACV = squared quantities in the C-VM formula
c           SACV = sum of the ACV values
c            CVM = values of the modified C-VM test statistic
c          ------------------------------------------------------
c           ablu = BLUE of location parameter a
c           bblu = BLUE of scale parameter b
c              c = null-hypothesized Pareto shape parameter
c            obs = number of observations in each of 5 cells
c          rtend = right endpoint of a cell
c             X2 = array of values of the Chi-square test statistic
c
c=================================================================
c
```

```
c  Input:
c          n = sample size = 5, 15, or 25 (from MAIN DO Loop 70)
c          P = array of n values of hypothesized CDF (from HYPCDF)
c       nshp = null-hyp shape counter (from MAIN DO Loop 90)
c       nalf = significance level counter (from MAIN DO Loop 80)
c       nsiz = sample size counter (from MAIN DO Loop 70)
c       nalt = alternate CDF counter (from MAIN DO Loop 60)
c       ablu = BLUE of location a (from BLCLE2 or BLCGT2)
c       bblu = BLUE of scale b (from BLCLE2 or BLCGT2)
c          c = hypothesized Pareto shape (from MAIN DO Loop 90)
c
c=====================================================================
c
c  Calculations for K-S test statistic (eqns 41 & 42):
c
c              DP(i) = ABS[ (i/n) - P(i) ]
c              DM(i) = ABS[ P(i) - (i-1)/n ]
c
c              DPLUS = max [ DP(i) ]  for i=1,2,...,n
c             DMINUS = max [ DM(i) ]  for i=1,2,...,n
c
c                 KS = max (DPLUS,DMINUS)
c
c--------------------------------------------------------------------
c
c  Calculations for A-D test statistic (eqn 43):
c
c              AL(j) = ln (P(j))
c              AM(j) = ln (1 - P(n+1-j))
c              AN(j) = AL(j) + AM(j)
c
c             AAA(j) = (2*j - 1) * AN(j)
c               SAAA = AAA(1) + AAA(2) + ... + AAA(n)
c
c                 AD = -n - (1/n) * SAAA
c
c--------------------------------------------------------------------
c
c  Calculations for C-VM test statistic (eqn 44):
c
c             ACV(k) = [ P(k) - (2*k - 1)/(2*n) ]**2
c               SACV = ACV(1) + ACV(2) + ... + ACV(n)
c
c                CVM = (1/(12*n)) + SACV
c
c--------------------------------------------------------------------
c
c  Calculations for Chi-square test statistic (eqn 62):
c
c         rtend(i) = ablu - bblu + bblu * (1 - .2*i) ** (-1/c)
c              ex = n / 5.
c
```

```
c       X2 = [(obs(1)-ex)**2] / ex   +  [(obs(2)-ex)**2] / ex
c            + ... + [(obs(5)-ex)**2] / ex
c
c=====================================================================
c
c  Declare Variables:
c
        common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1          Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,nrep,
     1          nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
        integer n,nsiz,nshp,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1          nrCV(2,2,3,8),obs(5),nrX2(2,2,3,8)
        real    x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1          CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1          P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1          CVpwr(2,2,3,8),DP(25),DM(25),DPLUS,DMINUS,AL(25),
     1          AM(25),AN(25),AAA(25),SAAA,ACV(25),SACV,rtend(4),
     1          X2crit(2,2,3),X2(2,2,3,8),ex
        double precision dseed
c
c -----   Compute the K-S Test Statistic (eqns 41 & 42):  -----
c
            DPLUS = 0
            DMINUS = 0
            do 5 ik = 1,25
               DP(ik) = 0
               DM(ik) = 0
  5         continue
c
            do 10 i = 1,n
c
              DP(i) = ABS( (i/real(n)) - P(i) )
              DM(i) = ABS( P(i) - (i-1)/real(n) )
c
c                 if (nshp.eq.1 .and. nalf.eq.2 .and. n.eq.5 .and.
c     1               nalt .lt. 3) then
c                 print*,'P(i)=',P(i),'DP(i)=',DP(i),'DM(i)=',DM(i)
c                 end if
c
  10        continue
c
        DPLUS =  MAX( DP(1),DP(2),DP(3),DP(4),DP(5),DP(6),DP(7),
     1             DP(8),DP(9),DP(10),DP(11),DP(12),DP(13),DP(14),
     1             DP(15),DP(16),DP(17),DP(18),DP(19),DP(20),
     1             DP(21),DP(22),DP(23),DP(24),DP(25) )
c
        DMINUS = MAX( DM(1),DM(2),DM(3),DM(4),DM(5),DM(6),DM(7),
     1             DM(8),DM(9),DM(10),DM(11),DM(12),DM(13),DM(14),
     1             DM(15),DM(16),DM(17),DM(18),DM(19),DM(20),
     1             DM(21),DM(22),DM(23),DM(24),DM(25) )
c
        KS(nshp,nalf,nsiz,nalt) = MAX(DPLUS,DMINUS)
```

```
c
c        print*,' '
c        print*,'* * * * * * * * * * * * * * * * * * * * * * * * *'
c        print*,' '
c        print*,'KS VALUES FROM TESTAT  -- ITERATION =',it
c        print*,'c=',c,'nalf=',nalf,'  **  n=',n,'  **  nalt=',nalt
c        print*,'KS Stat=',KS(nshp,nalf,nsiz,nalt),
c    1          '  ** DPLUS=',DPLUS,'  **  DMINUS=',DMINUS
c        print*,' '
c
c -------   Compute the A-D Test Statistic (eqn 43):   -------
c
        SAAA = 0
c
        do 20 j = 1,n
            AL(j) = log (P(j))
            AM(j) = log (1.0 - P(n+1-j))
            AN(j) = AL(j) + AM(j)
            AAA(j) = (2.0*j - 1.0) * AN(j)
            SAAA = SAAA + AAA(j)
  20    continue
c
        AD(nshp,nalf,nsiz,nalt) = -n - (1.0/real(n)) * SAAA
c
c -------   Compute the C-VM Test Statistic (eqn 44):   -------
c
        SACV = 0
c
        do 30 k = 1,n
            ACV(k) = ( P(k) - (2.0*k - 1.0)/(2.0*real(n)) )**2
            SACV = SACV + ACV(k)
  30    continue
c
        CVM(nshp,nalf,nsiz,nalt) = SACV + (1.0/(12.0*real(n)))
c
c -----  Compute the Chi-Square Test Statistic (eqn 62): -----
c
          do 40 in = 1,5
             obs(in) = 0
  40      continue
c
          do 50 ki = 1,4
             rtend(ki) = ablu-bblu + bblu*(1.-.2*ki)**(-1./c)
  50      continue
c
          do 60 m = 1,n
c
             if( x(m) .le. rtend(1) ) then
                obs(1) = obs(1) + 1
             else if (x(m).le.rtend(2)) then
                obs(2) = obs(2) + 1
             else if (x(m).le.rtend(3)) then
```

```
                 obs(3) = obs(3) + 1
              else if (x(m).le.rtend(4)) then
                 obs(4) = obs(4) + 1
              else
                 obs(5) = obs(5) + 1
              end if
c
 60        continue
c
           ex = n / 5.
c
           X2(nshp,nalf,nsiz,nalt) = ( (obs(1)-ex) **2 ) / ex
     1       + ((obs(2)-ex)**2)/ex + ((obs(3)-ex)**2)/ex
     1       + ((obs(4)-ex)**2)/ex + ((obs(5)-ex)**2)/ex
c
c         print*,' '
c         print*,'+ + + + +  + + + + + + + + + + + + + + + +'
c         print*,' '
c         print*,'X2 VALUES FROM TESTAT  -- ITERATION =',it
c         print*,'c=',c,'nalf=',nalf,'  **  n=',n,'  **  nalt=',nalt
c         print*,'RT ENDPOINTS OF INTERVALS:'
c         print*,rtend(1),rtend(2),rtend(3),rtend(4)
c         print*,'x(1)=',x(1),'x(10)=',x(10),'x(25)=',x(25)
c         print*,'OBSERVATIONS PER CELL:'
c         print*,'Cell 1:',obs(1),'  **  Cell 2:',obs(2)
c         print*,'Cell 3:',obs(3),'  **  Cell 4:',obs(4)
c         print*,'Cell 5:',obs(5)
c         print*,'CHI SQUARE TEST STAT:'
c         print*,'X2 Stat=',X2(nshp,nalf,nsiz,nalt)
c         print*,' '
c
        return
        end
c
c==================================================================
c                           END SUBROUTINE TESTAT
c******************************************************************
```

```
      Subroutine COMPAR
c*********************************************************************
c**                                                                 **
c**        B E G I N   S U B R O U T I N E   C O M P A R           **
c**                                                                 **
c*********************************************************************
c
c  Ref: Appendix B, Figure 7, Step 3.
c
c=====================================================================
c
c  Purpose:
c
c       Compare a test statistic, calculated from Chi-square or the
c       modified Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D),
c       or Cramer-von Mises (C-VM) test, against the appropriate
c       critical value. From a series of test statistics, count the
c       number of times the null hypothesis is rejected, i.e., the
c       number of test statistic values that exceed the critical
c       value.  The K-S, A-D, and C-VM critical values were taken
c       from Tables VI- VIII of the thesis.
c
c=====================================================================
c
c  Variables:
c              c = null-hypothesis Pareto shape parameter
c          alpha = significance level
c              n = sample size
c           nshp = shape parameter counter (1: c=1.0; 2: c=3.5)
c           nalf = significance level counter (1: α=.05; 2: α=.01)
c           nsiz = sample size counter (1: n=5; 2: n=15; 3: n=25)
c             KS = array of modified K-S test statistics
c            CVM = array of modified C-VM test statistics
c             AD = array of modified A-D test statistics
c             X2 = array of Chi-square test statistics
c
c====================================================================
c
c  Input:
c              c = null-hyp shape parameter (from MAIN DO Loop 90)
c          alpha = significance level (from MAIN DO Loop 80)
c              n = sample size (from MAIN DO Loop 80)
c           nshp = shape parameter counter (from MAIN DO Loop 90)
c           nalf = significance level counter (MAIN DO Loop 80)
c           nsiz = sample size counter (from MAIN DO Loop 70)
c           nalt = alternate CDF counter (from MAIN DO Loop 60)
c             KS = array of K-S test statistics (from TESTAT)
c            CVM = array of C-VM test stats (from TESTAT)
c             AD = array of A-D test statistics (from TESTAT)
c      KScrit(nshp,nalf,nsiz) = K-S critical values (Table VI)
c      ADcrit(nshp,nalf,nsiz) = A-D critical values (Table VII)
c      CVcrit(nshp,nalf,nsiz) = CVM critical values (Table VIII)
```

```
c      X2crit(nshp,nalf,nsiz) = Chi-square critical values
c
c=================================================================
c
c  Calculations: none
c
c=================================================================
c
c  Output:
c
c    nrKS = number of times null hypothesis is rejected under K-S
c    nrAD = number of times null hypothesis is rejected under A-D
c    nrCV = number of times null hypothesis is rejected under CVM
c    nrX2 = number of times null hyp is rejected under Chi-square
c
c=================================================================
c
c  Declare Variables:
c
        common  dseed,x,n,c,nc,B,D,ablu,bblu,P,Bsum1,Bxsum1,
     1          Bxsum2,Bxsm2c,KS,AD,CVM,it,nsiz,nshp,nrep,
     1          nalt,nalf,nrKS,nrAD,nrCV,nrX2,X2
        integer n,nsiz,nshp,it,nrep,nrKS(2,2,3,8),nrAD(2,2,3,8),
     1          nrCV(2,2,3,8),nrX2(2,2,3,8)
        real    x(25),ablu,bblu,B(25),D,KS(2,2,3,8),AD(2,2,3,8),
     1          CVM(2,2,3,8),c,nc,Bsum1,Bxsum1,Bxsum2,Bxsm2c,
     1          P(25),r(25),alpha,KSpwr(2,2,3,8),ADpwr(2,2,3,8),
     1          CVpwr(2,2,3,8),KScrit(2,2,3),ADcrit(2,2,3),
     1          CVcrit(2,2,3),X2crit(2,2,3),X2(2,2,3,8)
        double precision dseed
c
c         print*,'******************************************'
c              print*,'Numbers of Rejects at COMPAR Entrance'
c              print*,'c =',c,'nalf =',nalf,'n=',n,'nalt=',nalt
c              print*,'KS Rejects = ',nrKS(nshp,nalf,nsiz,nalt)
c              print*,'AD Rejects = ',nrAD(nshp,nalf,nsiz,nalt)
c              print*,'CV Rejects = ',nrCV(nshp,nalf,nsiz,nalt)
c              print*,'========================='
c
c  ---  Input K-S Critical Values from Table VI: ---
c
        KScrit(1,1,1) = .3676251
        KScrit(1,1,2) = .2157919
        KScrit(1,1,3) = .1698559
        KScrit(1,2,1) = .4074441
        KScrit(1,2,2) = .2468265
        KScrit(1,2,3) = .2007451
        KScrit(2,1,1) = .3493998
        KScrit(2,1,2) = .2376525
        KScrit(2,1,3) = .1886063
        KScrit(2,2,1) = .3815996
        KScrit(2,2,2) = .2743093
```

```
      KScrit(2,2,3) = .2182668
c
c --- Input A-D Critical Values from Table VII: ---
c
      ADcrit(1,1,1) = 1.236920
      ADcrit(1,1,2) =  .8907447
      ADcrit(1,1,3) =  .9147376
      ADcrit(1,2,1) = 2.076011
      ADcrit(1,2,2) = 1.250242
      ADcrit(1,2,3) = 1.311781
      ADcrit(2,1,1) =  .6840515
      ADcrit(2,1,2) =  .8985860
      ADcrit(2,1,3) =  .9520599
      ADcrit(2,2,1) =  .9126385
      ADcrit(2,2,2) = 1.268849
      ADcrit(2,2,3) = 1.449695
c
c --- Input C-VM Critical Values from Table VIII: ---
c
      CVcrit(1,1,1) = .1389776
      CVcrit(1,1,2) = .1312229
      CVcrit(1,1,3) = .1386932
      CVcrit(1,2,1) = .1738497
      CVcrit(1,2,2) = .1923594
      CVcrit(1,2,3) = .1988135
      CVcrit(2,1,1) = .1186844
      CVcrit(2,1,2) = .1561372
      CVcrit(2,1,3) = .1618638
      CVcrit(2,2,1) = .1574178
      CVcrit(2,2,2) = .2217665
      CVcrit(2,2,3) = .2403474
c
c --- Input Chi-square Critical Values : ---
c
      X2crit(1,1,1) = 6.000003
      X2crit(1,1,2) = 7.333337
      X2crit(1,1,3) = 7.600005
      X2crit(1,2,1) = 12.00000
      X2crit(1,2,2) = 10.66667
      X2crit(1,2,3) = 10.80000
      X2crit(2,1,1) = 6.000003
      X2crit(2,1,2) = 7.333337
      X2crit(2,1,3) = 7.600005
      X2crit(2,2,1) = 6.000003
      X2crit(2,2,2) = 10.46378
      X2crit(2,2,3) = 10.80000
c
c --- Compare Test Statistics vs Critical Values: ---
c
c     print*,'$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$'
c     print*,'BEFORE REJ COUNTER IS INCREMENTED:'
c     print*,'c =',c,'nalf =',nalf,'  **  n=',n,'  **  nalt=',nalt
```

```
c
c        print*,'KS Stat =',KS(nshp,nalf,nsiz,nalt),
c     1          ' Crit =',KScrit(nshp,nalf,nsiz)
c
c        print*,'AD Stat =',AD(nshp,nalf,nsiz,nalt),
c     1          ' Crit =',ADcrit(nshp,nalf,nsiz)
c
c        print*,'CV Stat =',CVM(nshp,nalf,nsiz,nalt),
c     1          ' Crit =',CVcrit(nshp,nalf,nsiz)
c
c        print*,'X2 Stat =',X2(nshp,nalf,nsiz,nalt),
c     1          ' Crit =',X2crit(nshp,nalf,nsiz)
c        print*,'$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$'
c
         if ( KS(nshp,nalf,nsiz,nalt) .gt. KScrit(nshp,nalf,nsiz) )
     1     nrKS(nshp,nalf,nsiz,nalt) = nrKS(nshp,nalf,nsiz,nalt) + 1
c
         if ( AD(nshp,nalf,nsiz,nalt) .gt. ADcrit(nshp,nalf,nsiz) )
     1     nrAD(nshp,nalf,nsiz,nalt) = nrAD(nshp,nalf,nsiz,nalt) + 1
c
         if (CVM(nshp,nalf,nsiz,nalt) .gt. CVcrit(nshp,nalf,nsiz) )
     1     nrCV(nshp,nalf,nsiz,nalt) = nrCV(nshp,nalf,nsiz,nalt) + 1
c
         if ( X2(nshp,nalf,nsiz,nalt) .gt. X2crit(nshp,nalf,nsiz) )
     1     nrX2(nshp,nalf,nsiz,nalt) = nrX2(nshp,nalf,nsiz,nalt) + 1
c
c            print*,'=========================================='
c            print*,'Numbers of Rejects at COMPAR Exit'
c            print*,'c =',c,'nalf =',nalf,' n=',n,' nalt=',nalt
c            print*,'KS Rejects = ',nrKS(nshp,nalf,nsiz,nalt)
c            print*,'AD Rejects = ',nrAD(nshp,nalf,nsiz,nalt)
c            print*,'CV Rejects = ',nrCV(nshp,nalf,nsiz,nalt)
c            print*,'X2 Rejects = ',nrX2(nshp,nalf,nsiz,nalt)
c            print*,'=========================================='
c
        return
        end
c
c=================================================================
c                        END SUBROUTINE COMPAR
c*****************************************************************
```

BIBLIOGRAPHY

1. Amstadter, B. Reliability Mathematics. New York: McGraw-Hill Book Company, 1971.

2. Anderson, T. W. and D. A. Darling. "Asymptotic Theory of Goodness of Fit Criteria Based on Stochastic Processes," Annals of Mathematical Statistics, 23: 193-212 (1952).

3. Anderson, T. W. and D. A. Darling. "A Test of Goodness of Fit," Journal of the American Statistical Association, 49: 765-769 (Dec 1954).

4. Andrews, D. F. and others. Robust Estimates of Location. Princeton University Press, 1972.

5. Banks, Jerry and John S. Carson. Discrete-Event System Simulation. Englewood Cliffs: Prentice-Hall, 1984.

6. Bell, C. B. and others. Signal Detection for Pareto Renewal Processes. Technical Report No. 8-82 for the Office of Naval Research. Contract N00014-80-C-0208. San Diego State University, San Diego CA, Oct 1982 (AD-A120 972).

7. Berger, J. M. and B. Mandelbrot. "A New Model for Error Clustering in Telephone Circuits," IBM Journal of Research and Development, 7: 224-236 (July 1963).

8. Brownlee, K. A. Statistical Theory and Methodology in Science and Engineering (Second Edition). New York: John Wiley and Sons, 1965.

9. Bush, J. G. and others. "Modified Cramer-von Mises and Anderson-Darling Tests for Weibull Distributions with Unknown Location and Scale Parameters," Communications in Statistics, Part A - Theory and Methods, 12: 240-245 (1983).

10. Buslenko, N. P., and others. The Monte Carlo Method. New York: Pergamon Press, 1966.

11. Champernowne, D. G. "The Graduation of Income Distributions," Econometrica, 20: 591-615 (1952).

12. Charek, Dennis J. A Comparison of Estimation Techniques for the Three-Parameter Pareto Distribution. MS Thesis, GSO/MA/85D-3. School of Engineering, Air Force Institute of Technology (AU), Wright Patterson AFB OH, December 1985.

13. Conover, W. J. Practical Nonparametric Statistics (Second Edition). New York: John Wiley and Sons, 1980.

14. David, F. N. and N. L. Johnson. "The Probability Integral Transformation When Parameters are Estimated from the Sample," Biometrika, 35: 182-190 (1948).

15. David, Herbert A. Order Statistics (Second Edition). New York: John Wiley and Sons, 1981.

16. Davis, Henry T. and Michael L. Feldstein. "The Generalized Pareto Law as a Model for Progressively Censored Survival Data," Biometrika, 66: 299-306 (1979).

17. Fisk, P. R. "The Graduation of Income Distributions," Econometrica, 29: 171-185 (1961).

18. Freiling, E. C. A Comparison of the Fallout Mass-Size Distributions Calculated by Lognormal and Power-Law Models. Report No. USNRDL-TR-1105 for the U.S. Naval Radiological Defense Laboratory, San Francisco CA, Nov 1966 (AD-646019).

19. Green, J. and Y. Hegazy. "Powerful Modified EDF Goodness-of-Fit Tests," Journal of the American Statistical Association, 71: 204-209 (1976).

20. Hajek, Jaroslav. A Course in Non-Parametric Statistics. San Francisco: Holden-Day, Inc., 1969.

21. Hammersley, J. M. and D. C. Handscomb. Monte Carlo Methods. London: Methuen and Co., 1967.

22. Harris, Carl M. "The Pareto Distribution as a Queue Service Discipline," Operations Research, 16: 307-313 (Jan-Feb 1968).

23. Harter, H. L. Order Statistics and Their Use in Testing and Estimation, Vol 2. Aerospace Research Laboratories, Wright-Patterson AFB OH, 1969.

24. Harter, H. L. "Another Look at Plotting Positions," Communications in Statistics, A13(13): 1613-1633 (1984).

25. Harter, H. L. "A Monte Carlo Study of Plotting Positions," Communications in Statistics, B14(2): 317-343 (1985).

26. Hastings, N. A. J. and J. B. Peacock. Statistical Distributions. London: Butterworth & Co. Ltd., 1974.

27. Hines, William W. and Douglas C. Montgomery. Probability and Statistics in Engineering and Management Science. New York: The Ronald Press Co., 1972.

28. Johnson, Norman L. and Samuel Kotz. Continuous Univariate Distributions-1. Boston: Houghton Mifflin Co., 1970.

29. Kaminsky, Kenneth S. Best Linear Unbiased Prediction of Order Statistics in Exponential and Pareto Populations. Contract F33615-71-C-1463. Technical Report No. ARL 75-0201 for Aerospace Research Laboratories, Wright-Patterson AFB OH, June 1975 (AD-A014 740).

30. Kaminsky, Kenneth S. and Paul I. Nelson. "Best Linear Unbiased Prediction of Order Statistics in Location and Scale Families," Journal of the American Statistical Association, 70: 145-150 (1975).

31. Kapur, K. C. and L. R. Lamberson. Reliability in Engineering Design. New York: John Wiley and Sons, 1977.

32. Koutrouvelis, Ioannis. Estimation of Asymptotic Pareto Laws and the Tail of a Distribution. Contract Number N00014-72-C-0508. Technical Report No. 34 for Office of Naval Research, Arlington VA, Aug 1975 (AD-A018 173).

33. Kulldorff, Gunnar and Kerstin Vannman. "Estimation of the Location and Scale Parameters of a Pareto Distribution by Linear Functions of Order Statistics", Journal of the American Statistical Association, 68: 218-227 (1973).

34. Lilliefors, H. "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown", Journal of the American Statistical Association, 62: 399-402 (1967).

35. Lilliefors, H. "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown", Journal of the American Statistical Association, 64: 387-399 (1969).

36. Littel, Ramon C., James McClave, and Walter Offen. "Goodness-of-Fit Tests for the Two Parameter Weibull Distribution", Communications in Statistics, B8(3): 257-269 (1979).

37. Little, Robert E. Probability and Statistics for Engineers. Champaign IL: Matrix Publishers, Inc., 1978.

38. Mann, N. R., E. M. Scheuer, and K. W. Fertig. "A New Goodness-of-Fit Test for the Two-Parameter Weibull or Extreme-Value Distribution with Unknown Parameters", Communications in Statistics, 2: 383-400 (1973).

39. Massey, Frank J. "The Kolmogorov-Smirnov Test for Goodness of Fit", Journal of the American Statistical Association, 46: 68-78 (1951).

40. Mood, A. M. and F. A. Graybill, Introduction to the Theory of Statistics (Second Edition). New York: McGraw Hill Inc., 1963.

41. Moore, Albert H. and H. L. Harter. "One-order-statistic Conditional Estimators of Shape Parameters of Limited and Pareto Distributions and Scale Parameters of Type II Asymptotic Distributions of Smallest and Largest Values," IEEE Transactions on Reliability, R-16: 100-103 (1967).

42. Pigou, A. C., The Economics of Welfare. London: Macmillan and Co., 1948.

43. Ream, Thomas J. A New Goodness of Fit Test for Normality with Mean and Variance Unknown. MS Thesis, GOR/MA/81D-9. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, Dec 1981.

44. Steindl, Josef. Random Processes and the Growth of Firms. New York: Hafner Publishing Co., 1965.

45. Stephens, M. A. "EDF Statistics for Goodness of Fit and Some Comparisons", Journal of the American Statistical Association, 69: 730-737 (Sep 1974).

46. Stephens, M. A. "Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters", Annals of Statistics, 4: 357-369 (1976).

47. Stephens, M. A. The Anderson-Darling Statistic. Grant No. DAAG29-77-G-0031. Technical Report No. 39 for the U.S. Army Research Office, Dept. of Statistics, Stanford University, Stanford CA, Oct 1979 (AD-A079 807).

48. Vannman, Kerstin. "Estimators Based on Order Statistics from a Pareto Distribution", Journal of the American Statistical Association, 71: 704-708 (Sep 1976).

49. Viviano, Philip J. A Modified Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises Test for the Gamma Distribution with Unknown Location and Scale Parameters. MS Thesis, GOR/MA/82D-4. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, Dec 1982.

50. Wingo, Dallas R. "Estimation in a Pareto Distribution: Theory and Computation", IEEE Transactions on Reliability, R-28: 35-37 (Apr 1979).

51. Wong, Wing-Yue. On the Property of Dullness of Pareto Distribution. Contract No. N00014-75-C0455. Technical Report No. 82-16 for the Office of Naval Research, Purdue University, West Lafayette IN, May 1982 (AD-A119 631).

52. Woodbury, Larry B. A New Goodness of Fit Test for the Uniform Distribution with Unspecified Parameters. MS Thesis, GOR/MA/82D-6. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, Dec 1982.

53. Woodruff, Brian W. and others. "A Modified Kolmogorov-Smirnov Test for Weibull Distributions with Unknown Location and Scale Parameters," IEEE Transactions on Reliability, R-32: 209-213 (Jun 1983).

54. Yoder, John D. Modified Kolmogorov-Smirnov, Anderson-Darling, and Cramer-Von Mises Tests for the Logistic Distribution with Unknown Location and Scale Parameters. MS Thesis, GOR/ENC/83D. School of Engineering, Air Force Institute of Technology (AU), Wright Patterson AFB OH, December 1983.

VITA

Captain James E. Porter III was born in Tokyo, Japan, on 24 September 1951. He graduated from Judson High School, Converse, Texas, in 1969. He then attended the University of Texas at Austin and in 1974 graduated Phi Beta Kappa with a Bachelor of Science degree in Mathematics.

Upon completing Officer Training School and receiving his USAF commission in April 1975, he was assigned to the Space Systems (now called Space Operations) career field, Air Force Specialty Code (AFSC) 20XX. He served as a Space Surveillance Officer at the Sea-Launched Ballistic Missile Detection and Warning radar site, Fort Fisher AFS, North Carolina, from June 1975 to May 1977; and at the Ballistic Missile Early Warning System radar site, Thule, Greenland, from May 1977 to May 1978.

From June 1978 to May 1981 Captain Porter was assigned to Headquarters North American Aerospace Defense Command, Peterson AFB, Colorado, as a Space Systems Staff Officer. He next served as Space Operations Career Management Staff Officer, Air Force Manpower and Personnel Center, Randolph AFB, Texas, until May 1984. He then entered the Graduate Space Operations Program, School of Engineering, Air Force Institute of Technology.

Address: 4026 Kirby Drive, San Antonio, Texas 78219.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A163 837

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): AFIT/GSO/MA/85D-6	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: School of Engineering; 6b. OFFICE SYMBOL (If applicable): AFIT/ENS	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State and ZIP Code): Air Force Institute of Technology, Wright-Patterson AFB OH 45433-6583	7b. ADDRESS (City, State and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State and ZIP Code)	10. SOURCE OF FUNDING NOS.: PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT NO.
11. TITLE (Include Security Classification): See Box 19	

12. PERSONAL AUTHOR(S): James E. Porter III, Captain, USAF

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
MS Thesis	FROM ____ TO ____	1985 December	183

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	Monte Carlo Method; Statistical Functions; Probability Distribution Function; Statistical Analysis; Statistical Decision Theory; Statistical Tests; Order Statistics
12	01		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

TITLE: *MODIFIED KOLMOGOROV-SMIRNOV, ANDERSON-DARLING, AND CRAMER-VON MISES TESTS FOR THE PARETO DISTRIBUTION WITH UNKNOWN LOCATION AND SCALE PARAMETERS*

THESIS ADVISOR: Dr Albert H. Moore
Professor of Mathematics

Approved for public release: IAW AFR 190-1.
Lyn Wolaver 16 JAN 86
LYNN E. WOLAVER
Dean for Research and Professional Development
Air Force Institute of Technology (ATC)
Wright-Patterson AFB OH 45433

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL: Prof. Albert H. Moore	22b. TELEPHONE NUMBER (Include Area Code): (513) 255-3098; 22c. OFFICE SYMBOL: AFIT/ENC

DD FORM 1473, 83 APR — EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

19. ABSTRACT

Modified Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Cramer-von Mises (C-VM) critical values are generated for the three-parameter Pareto distribution. The values may be used to test whether a set of observations follows a Pareto distribution when the location and scale parameters are unspecified and thus must be estimated from the sample. A Monte Carlo simulation of 5000 repetitions is used to generate critical values for sample sizes 5(5)30 (i.e., 5 to 30 in increments of 5) and Pareto shape parameters .5(.5)4.0.

A 5000-repetition Monte Carlo investigation is carried out by using 5, 15, and 25 observations from eight alternate distributions to compare the powers of the K-S, A-D, C-VM, and Chi-square tests. The power values of the tests are relatively low for a sample size of five. However, the powers of the modified K-S, A-D, and C-VM tests are considerably better than the Chi-square test at larger sample sizes. Next to the Chi-square test, the A-D test has the lowest power in most cases.

A functional relationship is identified between the modified K-S and C-VM test statistics and the Pareto shape parameter. The critical values are found to be a linear function of the shape parameters between 1.5 and 4.0.

END

FILMED

3-86

DTIC